Lights, Camera, Action! Strategies to Create a Blockbuster Movie

MGT 6203 Team 81 Final Report – Spring 2023

Zach Geesing, Raymond Walther, Christin Whitton, Kevin Fu, Fendy Setiawan

1. Introduction

1.1 Background Information

American novelist, playwright and screenwriter William Goldman once said, "*Nobody knows anything.... Not one person in the entire motion picture field knows for a certainty what's going to work. Every time out it's a guess and, if you're lucky, an educated one.*" There are many elements to consider when making a movie, including star power, the overall budget, the director, the timing and method of movie release, length of the film, the advertising budget, whether the movie is an original or a sequel, and the movie rating. Understanding what typically works in film can help filmmakers to make educated decisions as they move forward with new projects. Identifying the factors that influence movie profitability will help companies (e.g. movie franchises or production companies) create strategies for making, promoting, and releasing the most profitable movies. Movie producers can advertise movies during months or years when certain genres are more popular or prioritize releases that may be more popular during periods of economic downturn. These strategies and educated decisions can translate directly into financial gain for those involved in this process.

1.2 Literature Review

In 1983, Barry Litman embarked on one of the first analyses of movies, with a similar goal as ours - to quantify success in the film industry. He used multiple regression analysis. Using theater rentals as his dependent variable, he found that production costs, critical ratings, certain genres (science fiction and horror), what kind of company distributed the film, whether it was released during the Christmas season and whether it was nominated for and/or won an academy award were significant. Three decades later, Pangarker and Smith from University of Stellenbosch Business School performed similar analyses to Litman, encompassing just the years 2009 and 2010, including films that were released on the international market. They found that the production cost, whether the movie was released by a major studio, if the film was a sequel, and if the film was nominated for an award, were all significant variables.

In 2015, Lindner, Lindquist and Arnold examined the processes that contribute to the under-representation of women in film by linking the impact on box office returns to depiction of gender. They hypothesized that due to the deeply socialized ranking of men over women, lower box office returns for movies with an independent female presence are likely to be the product of consumer demand, which is a downstream effect. They found that movies with large production budgets tend not to have a strong female presence, so movies with an independent female presence tend to have a smaller production budget. This appears to be the consequence of institutional upstream effects rather than the product of consumer demand.

1.3 Problem Statement and Hypothesis

The purpose of our investigation and analysis is to determine the key factors that contribute to making and marketing a profitable movie. Our initial hypothesis was that high-grossing movies were made by top production companies, a continuation of a successful movie series, directed by a well-known director, and starred a famous lead actor or actress. We also believed that genre, lead cast and director gender, release month and runtime might have some impact on the profitability of a movie, but were unsure of the specific influence. Finally, we anticipated that external factors, such as general economic conditions and the price of a movie ticket at the time of release, might also have some impact on movie profitability.

Based on our problem statement and initial hypotheses, we identified the following potential predictor variables for our analysis:

- Independent variables (predictor), sorted by hypothesized importance: production company, original/sequel, main cast popularity, director popularity, main cast gender, director gender, budget, genre, release month, and runtime.
- Additional independent variables: monthly inflation rate, unemployment rate, median household income, and movie ticket price.
- Dependent variable: movie profit (inflation-adjusted).

We also hypothesized that the significant factors might change over time, while some others would likely remain relevant throughout the decades. To better observe this, we split our dataset into four decades (1980s, 1990s, 2000s, and 2010s). These four decades were selected considering the amount of available data, since the completeness of more recent data would increase the chance of obtaining meaningful and relevant results.

2. Data Preparation

2.1 Overview of Datasets

Our primary dataset, the TMDB Movies Dataset obtained from Kaggle, contains details on over 700,000 movies over the past 100 years, including movie titles, release dates, production companies, genres, runtimes, and revenues. Additional key variables, such as cast, director, and production country were populated from the original TMDB dataset queried using an API key. Additionally, the API was used to impute data which was missing from the Kaggle dataset.

Economic indicators used for this study include average monthly Core Consumer Price Index (CPI), monthly historical unemployment rate, and annual household income. CPI was used as a measure of inflation, as well as to adjust financial-related variables (budget, revenue, and profit) to the same basis year (i.e. 2023 dollar). Historical unemployment rate is commonly used as a key indicator to gauge economic conditions, such as growth or recession, and was drawn from the Bureau of Labor Statistics (BLS). Median U.S. household income from 1980 to 2019 was obtained from the US Census Bureau. Unlike CPI and unemployment rate, household income is reported once a year. Historical U.S. movie ticket price, obtained from National Association of Theatre Owners, inflation-adjusted, was also included as one candidate predictor variable.

Data augmentation, filtering, and cleaning for this study was quite extensive due to the large number of potential predictor variables, and the fact that most of the data is text data (original language, genres, production companies, etc.), which required preprocessing before analyses could be performed. The next few subsections explain how we prepared and compiled the various datasets into a single cohesive dataset.

2.2 Data Augmentation

As mentioned in section 2.1, some key variables were not readily available in the Kaggle TMDB Movies Dataset. Therefore, we wrote a Python script to perform API calls on the original TMDB database to augment the existing, prefiltered movie dataset. Variables added during this process include genre, main cast popularity, director popularity, production companies, production country, and sequel data.

The process of querying data was limited to 40 queries per second by the TMDB API protocol which appeared to be enforced after about 10 seconds of continuously querying data. Three separate datasets were queried: (a) movie details for general movie release information, (b) credits for cast and crew data, and (c) collections for movie sequel information. Datasets were joined using their primary and foreign IDs to obtain the variables used in this study. One particular binary variable 'is_sequel' required conditional analysis.

For instance, Star Wars: A New Hope (1977) was part of the Star Wars collection, but it is not a sequel because it was the earliest movie to be released within the collection.

2.3 Data Filtering

We were mostly interested in data in recent decades and movie data with little to no missing values on key variables. Therefore, we performed data filtering as follows:

- Filtered to include only movies with release dates between 1980 and 2019 inclusive.
- Filtered out movies not produced in the U.S., as this study focuses on US-produced movies.
- Removed movies with no revenue or budget information, which are needed to compute profit, the response variable for this study.
- Removed movies without a lead cast or a director (<1% of rows).
- Removed some variables, e.g. original language and adult (whether a movie is intended only for adult audience) which had little to no variability.
- Removed duplicate movies (based on ID).

After the initial filtering of the movie data, we ended up with 5,568 movies to be cleaned.

2.4 Data Cleaning & Merging

To get the most value out of our dataset in the analysis phase, we needed to combine our datasets into a single data frame, handle any remaining missing values, and create several additional variables to better utilize the text-based columns in the movie dataset (genres, production companies, etc.). We performed data cleaning and merging as follows prior to exploratory data analysis (EDA):

- Converted variables to appropriate data types (e.g. factor, integer, etc.) for analysis.
- Imputed/removed missing data: missing genres and production companies imputed with "unknown" (approximately 1% of rows).
- Created new variables to identify release decade, release year, and release month of each movie. This was used to group the movies into the four separate decades and join the movie data with the additional economic datasets on release month and year.
- Calculated profit as revenue minus budget.
- Identified the top 5 production companies within each decade based on number of releases and created an indicator variable to identify movies produced by a top 5 company.
- Transformed genre information into one-hot encoded variables, with a total of 20 unique genres and 1 unknown genre.
- Calculated monthly inflation from cumulative CPI data.
- Converted all financial variables (e.g. profit, budget, household income, etc.) to 2023-equivalent dollars using CPI.
- Created adj_profit_2023 variable which transforms all profit values to a minimum of 1 so that logarithmic transformation may be performed on profit.

After performing the data cleaning process, we noticed a subset of movies with small budgets that we deemed unrealistic, as well as movies with very short runtimes. After looking into the data, we determined that they appeared to be short films and not actually full movies. Therefore, we performed additional data filtering by removing movies with budget or revenue below $10,000 or having a runtime below 60 minutes. In the end, we were left with 5,402 movies to be analyzed. Figure 1 shows budget data after filtering. The data was still skewed left, but we were more confident with the quality of the data points.

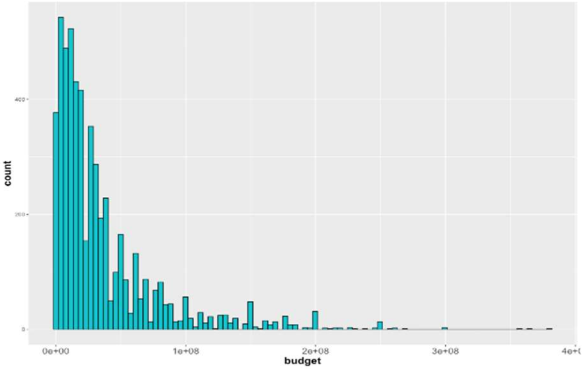


Figure 1. Histogram of budget adjusted to 2023 dollars

3. Exploratory Data Analysis (EDA)

Once we completed the data cleaning process, we performed EDA to better understand any patterns or correlations in our dataset and identify potential trends across the four decades in our study. Figure 2 shows the total number of movies in our dataset by decade, with the 2000s and 2010s containing the majority of our datapoints. Figure 3 shows distribution of profit by decade, the average profits across the decades were similar with the 2010s having slightly higher average profit.

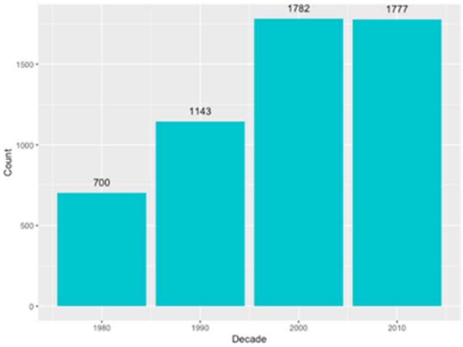


Figure 2. Bar Plot of Total Movies by Decade

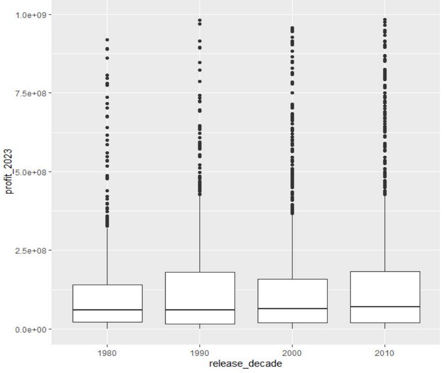


Figure 3. Boxplot of Profit by Decade

By exploring individual factors' correlation with profit, we found that some such as movie ticket price, unemployment rate, and director popularity did not appear to have a clear correlation with higher movie profits like we expected. However, we identified several factors outlined below that did show positive correlation with profitability.

Timing of movie release appears to have a positive impact on profitability, with those released in the summer (May to July) or towards the end of the year during the holiday season (November to December) showing the highest average profits, while those released during the first few months of the year show lower profits (Figure 4).

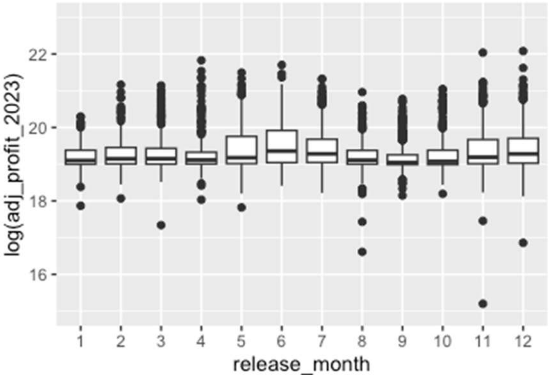


Figure 4. Boxplot of Movie Profit by Release Month

As we hypothesized, certain genres such as Action, Adventure and Animation are correlated with higher profit than other genres (Figure 5). Some of these higher-profit genres also have movies that appear to be potential outliers but actually turned out to be extremely high-grossing films such as Avatar, Titanic and ET.

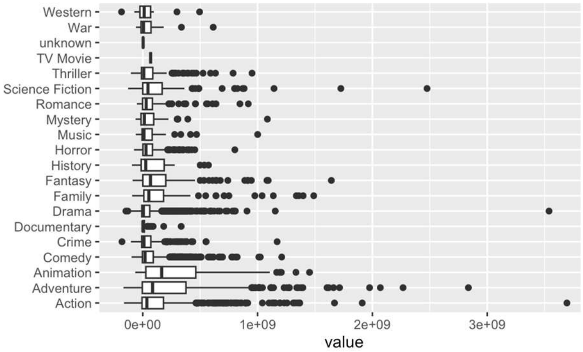


Figure 5. Movie Profit Distribution by Genre

One of our primary goals was to identify the impact of gender on movie profit (lead cast / director). During our exploration, we found that although movie profitability does seem to be increasing for female-led roles in some decades, profitability is still slightly higher for male vs female leads within each decade (Figure 6).

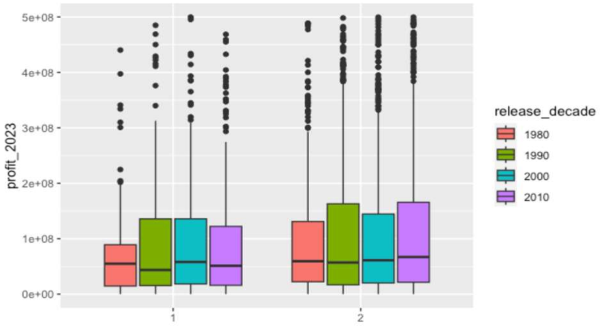


Figure 6. Movie Profit Distribution by Lead Cast Gender and Decade (female = 1, male = 2)

Other individual factors that showed positive correlation with profit included whether a movie was a sequel (higher profit for sequels), and movie budget. We also found that being produced by a top 5 production company seems to have a positive correlation with profit. Figure 7 shows profit distribution for sequel vs non-sequel movies, and Figure 8 shows the correlation between movie budget and adjusted profit. Figure 9 shows the correlation between profit and whether a movie is produced by a top 5 production company.

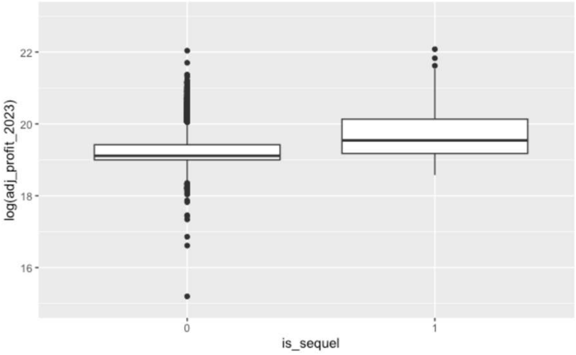


Figure 7. Profitability for Sequel (1) vs Non-sequel (0)

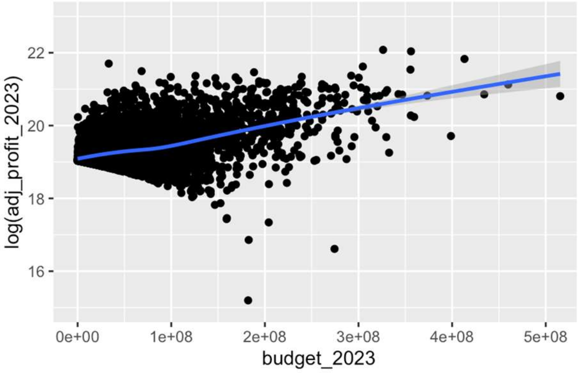


Figure 8. Movie Budget vs Profit

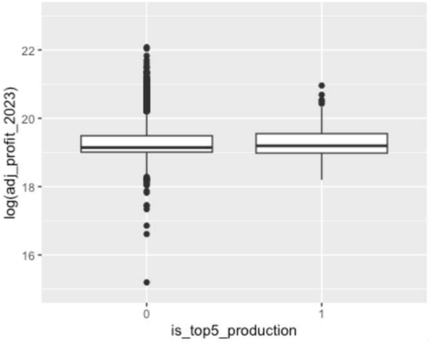


Figure 9. Top 5 Production Company Y/N vs Profit

4. Analysis Methodology

Two primary types of analyses were selected. First, linear regression was used for descriptive/ predictive analytics to determine which factors contribute to movie profitability. We also performed decision tree analyses to identify the combination of factors at the various nodes that lead to the highest profitability, which could be used in more of a prescriptive way.

To identify the relevant variables that have the largest influence on profitability of a movie, we segmented the full dataset into discrete decades. With these four separate decades of movies, we performed both linear regression and decision tree analysis to distill which features these different modeling approaches determine most relevant for predicting movie profitability. This allowed us to identify potential correlations between movie profit and movie-specific attributes as well as any relevant macroeconomic factors within each time period.

After completing the modeling for each decade, we analyzed the results using p-values to identify the statistically significant features, and how these differ from one decade to another. We also determined which features remained relevant across all decades, as we would consider these as being relevant today, and vital to our strategy of producing a profitable movie in the future.

4.1 Linear Regression

For our linear regression analysis, we started with all the variables across the decades, with the top production companies as a binary variable, as well as the top genres as a binary variable. Using stepwise regression, we were able to eliminate some variables with p-values > 0.05 (including director gender) and ran the model with and without the economic factors, to determine how important they are in predicting profitability.

We found that one-hot encoding the genres led to a better and more informative fit, and better accounted for movies that fell within multiple genres. We also performed regression on each individual variable to see whether the model could be improved further – we found that using the log of director popularity and lead cast popularity variables resulted in a more accurate model, as well as using polynomial regression for both the movie popularity and budget variables.

We ran our final model with and without economic variables across all decades, as well as within each individual decade, and compared the R^2 values of each model (see the results section for the table of R^2 values).

4.2 Decision Tree

Decision Trees were selected as one means of analyzing our dataset due to the clear interpretability of the results. The first node of a decision tree, also called the root node, is the node which best splits the data. In our case, when we included all possible features, our produced trees had root nodes of popularity as shown in Figure 10. This intuitively makes sense – we would expect movies with higher popularity scores to have been more profitable than those which were less popular. However, for our use case, it can be assumed that all movie makers are attempting to create the most popular movie possible, and as such it doesn't deliver much benefit to us in our decision tree analysis. Thus, we removed popularity and reran our decision tree regression, and the result is shown in Figure 11. This iteration had similar results with budget dominating not only the root node, but the first few nodes following the root. In short, the most profitable movies tend to have the highest budgets.

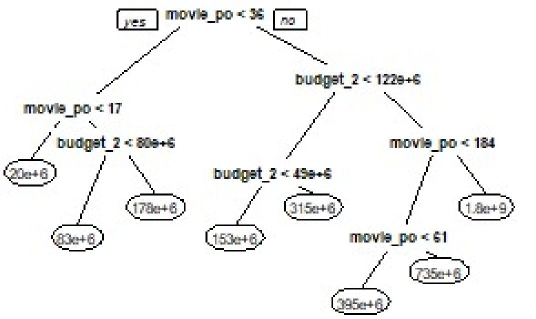


Figure 10. Decision Tree Model with All Predictors

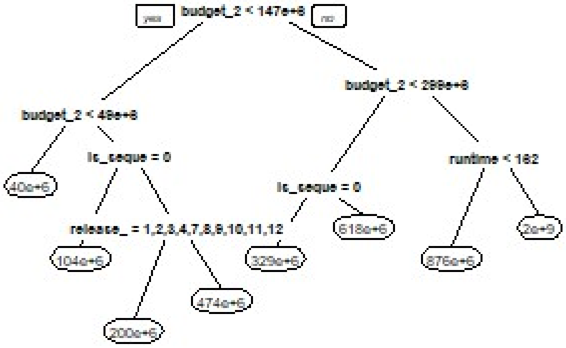


Figure 11. Decision Tree Model with movie_popularity Removed

Ultimately, due to the overwhelming dominance of these two features (popularity and budget) in our decision tree models, we reran our analysis with both highly influential features removed. After doing so, the decision trees offered better insight into the next most influential features. To avoid overfitting, the R^2 value and number of leaves were plotted to prune the tree using the "elbow method". An example using movies from the 1980s is shown in Figure 12. Initially, the tree model had 8 leaves with an R^2 value of 0.355; however, with only four leaves, we achieved an R^2 of 0.345.

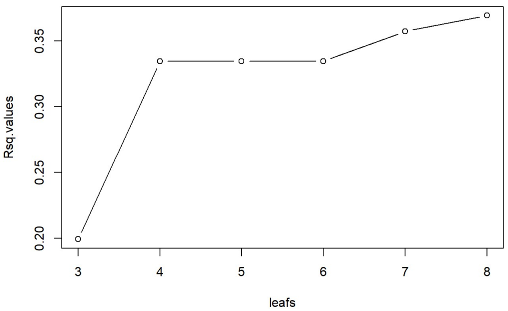


Figure 12. Demonstration of Pruning a Decision Tree

5. Analysis Results

5.1 Linear Regression

Our final equations for our linear regression were:

Without economic data:

Profit = β_0 + β_1 runtime + β_2 is sequel + β_3 log(lead cast popularity) + β_4 lead cast gender + β_5 log(director popularity) + β_6 release month + genres + β_{25} (movie popularity)^2 + β_{26} movie popularity + β_{27} (budget)^2 + β_{28} budget

With economic data:

Profit = β_0 + β_1 runtime + β_2 is sequel + β_3 log(lead cast popularity) + β_4 lead cast gender + β_5 log(director popularity) + β_6 release month + genres + β_{25} (movie popularity)^2 + β_{26} movie popularity + β_{27} (budget)^2 + β_{28} budget + β_{29} CPI monthly + β_{30} median income + β_{31} movie ticket price + β_{32} unemployment rate

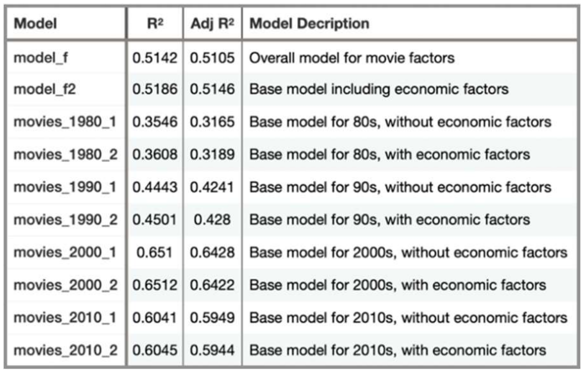
For both models, the genres variable is defined as:

genres = β_7 Action + β_8 Comedy + β_9 Crime + β_{10} Documentary + β_{11} Drama + β_{12} Family + β_{13} Fantasy + β_{14} History + β_{15} Horror + β_{16} Music + β_{17} Mystery + β_{18} Romance + β_{19} Science Fiction + β_{20} Thriller + β_{21} TV Movie + β_{22} War + β_{23} Western + β_{24} Unknown genre

We ran these models for each decade (omitting unknown and TV movie genres for the 1980s and unknown genre for the 1990s – there were no movies for those variables in those subsets), as well as with all the data. We found that the economic variables had minimal impact on the R^2 value, and one of the economic variables (CPI) showed up as significant only once, in our 1980's model.

We found that some variables were significant through the decades. These include runtime, director popularity, adventure as the genre and movie popularity. Other variables were shown to be significant in different decades – for instance, the month of release was not significant for our 1980's model but was found to be significant in our later models.

Table 1. R-squared values for various models



Model	R²	Adj R²	Model Decription
model_f	0.5142	0.5105	Overall model for movie factors
model_f2	0.5186	0.5146	Base model including economic factors
movies_1980_1	0.3546	0.3165	Base model for 80s, without economic factors
movies_1980_2	0.3608	0.3189	Base model for 80s, with economic factors
movies_1990_1	0.4443	0.4241	Base model for 90s, without economic factors
movies_1990_2	0.4501	0.428	Base model for 90s, with economic factors
movies_2000_1	0.651	0.6428	Base model for 2000s, without economic factors
movies_2000_2	0.6512	0.6422	Base model for 2000s, with economic factors
movies_2010_1	0.6041	0.5949	Base model for 2010s, without economic factors
movies_2010_2	0.6045	0.5944	Base model for 2010s, with economic factors

The coefficient for log(lead cast popularity) was mostly not significant, while log(director popularity) was always significant and positive, with a downward trend through the decades (Figure 13). May, June, July and December had the most positive coefficients (Figure 14).

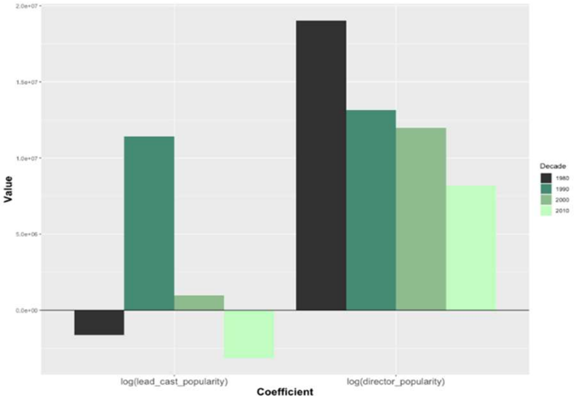


Figure 13. Cast and Director Popularity Coefficients

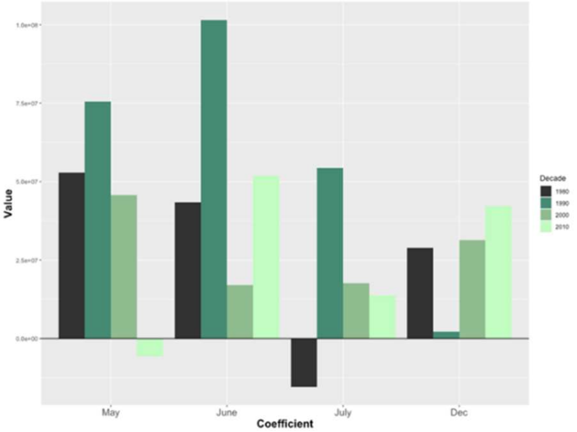


Figure 14. Top Release Month Coefficients

Genres have significant impact to movie profitability. The most profitable genres with positive coefficients are action, adventure, and animation. Adventure was significant and positive across all decades, while animation seemed to have a positive trend with positive and significant coefficients in later decades (Figure 15). The least profitable genres: crime, history and horror all had negative coefficients (Figure 16).

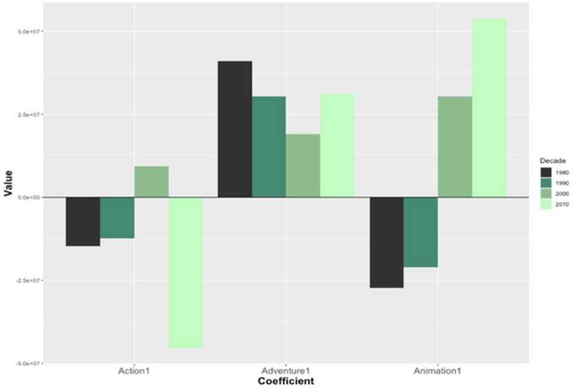


Figure 15. Most Profitable Genres Coefficients

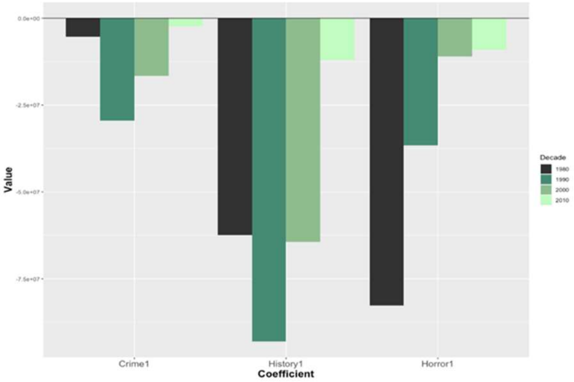


Figure 16. Least Profitable Genres Coefficients

5.2 Decision Tree

The variables included for our decision tree model are runtime, is sequel, lead cast popularity, lead cast gender, director gender, director popularity, release month, all hard-coded genres, CPI monthly, median income, movie ticket price, and unemployment rate.

We experimented with two types of decision tree models: classification tree (split profits into quintiles and deciles) and regression tree (inflation-adjusted profits). However, the results from the classification tree were overly simplified (as shown in Figure 17) and did not provide useful insights. Therefore, the following discussion will focus on the regression tree models.

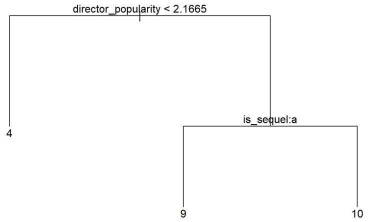


Figure 17. Classification Decision Tree Model

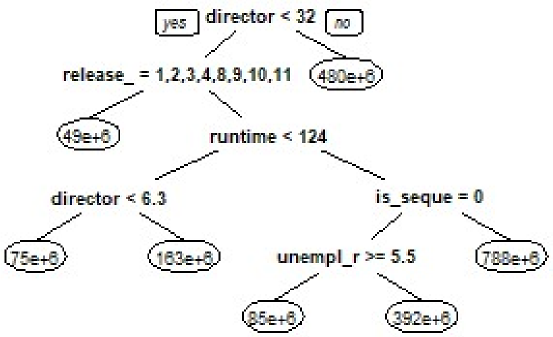


Figure 18. Decision Tree Model for 1980s Movies

For the 1980s, having a popular director and releasing a movie either during the summer months or in December prove to be the most influential features in our model (Figure 18).

Movies with the most popular directors on average earned approximately 5 times higher profit than movies with less popular directors. Next, the decision tree shows the time of year a movie released was very important – summertime and December releases were on average at least twice as profitable as those released in less optimal months. The next junction is movie runtime. This feature shows up in a few of our models, and they all had a positive impact on profits. One macroeconomic factor, the unemployment rate, had a noticeable impact on profits. Movies released during a low unemployment environment (< 5.5%) were over 4 times as profitable as those released during a high unemployment environment.

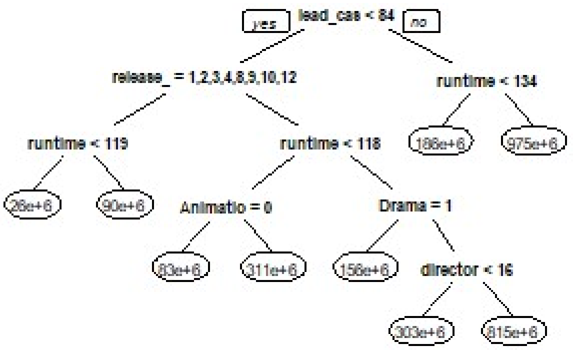


Figure 19. Decision Tree Model for 1990s Movies

For movies released in the 1990s (Figure 19), a long movie with a popular lead cast generated the highest profits on average. Unlike the 1980s, the peak month in winter is November instead of December. Genre also played an important role in the 1990s, animation movies generated higher profits and drama movies generated lower profits. Finally, for non-drama movies, popular directors could lead to almost 3 times higher profits.

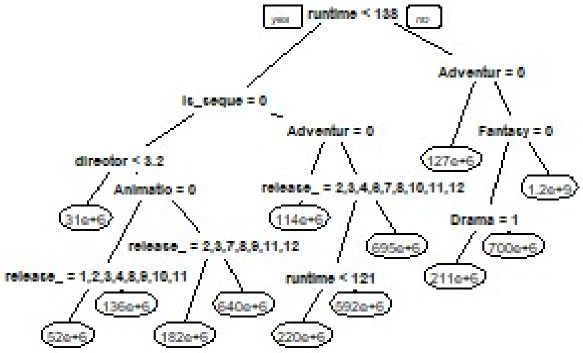


Figure 20. Decision Tree Model for 2000s Movies

Like the 1980s and 1990s, the runtime of movies released in the 2000s had a positive effect on profits (Figure 20). Genre is also an important factor. Adventure, fantasy, and animation movies all had greater profits on average, but drama had a negative impact on profits. Movies released during the summer also had higher profits on average.

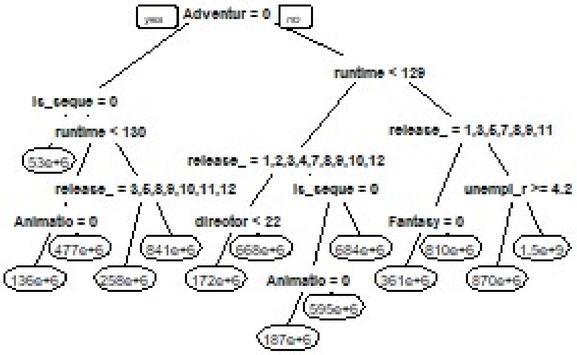


Figure 21. Decision Tree Model for 2010s Movies

Genres including adventure, animation, and fantasy had a positive effect on profitability for movies released in the 2010s, and adventure is the root node of the decision tree (Figure 21). Next, a sequel or long runtime can also lead to higher profits. Like other decades, movies released during certain months had a positive effect on profits. It is also interesting to see the effect of the unemployment rate at the bottom right of the tree. The lower unemployment rate resulted in about 70% higher profits.

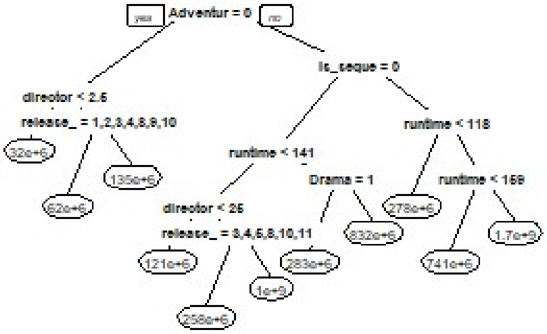


Figure 22. Decision Tree Model for Movies across Decades

Overall, adventure genre is the root node and had a positive effect on profits (Figure 22). Other factors leading to higher profits include a sequel, a popular director, a longer runtime, or release during summer and winter months.

6. Conclusion

Based on our linear regression models across the decades, we found that macroeconomic factors had little to no impact on the profitability of movies. This is not what we anticipated – it appears that this is not a significant factor to consider when determining how to create a profitable movie. We found that the sequel variable was highly significant in the 1980s, 1990s, 2010s, but not in the 2000s or in the model we created with all the decades. This indicates that movies that are sequels seem to increase profitability and are worth considering when evaluating projects.

Within each distinct decade, the variables found to be significant varied, as we expected. The variables significant across all decades were runtime, director popularity, genre, and movie popularity. Some of the notable variables whose significance changed with time were the genres: horror, crime, history and romance, and the release months of summertime (May to July) and holiday season (November to December).

Similar to the linear regression analyses, the decision tree models also showed that release months, runtime, sequel, and director popularity are consistently good indicators of highly profitable movies across the decades, which confirmed the hypothesis that release month, a continuation of a successful movie series, or movies directed by a well-known director led to higher profitability. However, the hypothesis that top production companies and popular lead actor/actress help in generating higher profits was not confirmed. Some factors leading to profitable movies have also changed throughout the decades. For example, the popularity of lead actor/actress is only important in the 1990s but not in other decades. The decision tree model can be a great complement to the linear regression model when strategizing planning and releasing a new movie.

7. Further Studies

This study attempted to find correlation between lead cast gender on movie profitability, which was not found to be statistically correlated. Lead cast gender is only one element of diversity. Further research is warranted to determine whether the gender diversity of an entire team of casts has any impact on movie profitability.

It is also important to further investigate whether there is another variable that can explain the correlation between a variable and the profitability of a movie. For example, whether blockbuster movies such as Avatar which was released in December was a big success because it was already highly anticipated before its release or because in general more people watch movies in December.

Another possible research question would be whether the unpredictability of movie plot (where the movie finale is not easily guessed) and emotions invoked while watching the movie (blockbuster movies tend to bring the audience into a roller-coaster of emotions) might impact movie profitability. The dataset for this study is not readily available and might require advanced deep learning models or expert opinions.

Budget allocation might also be important to study further. Allocating budget into the essential parts of a movie production might increase the quality of a movie and indirectly affect movie profit. However, budget data availability is limited and appears to be well-guarded by each production company.

8. References

Goldman, W. (1983). *Adventures in Screen Trade*. Grand Central Publishing.

Pangarker, N., & Smit, E. (2013). *The determinants of box office performance in the film industry revisited. South African Journal of Business Management*, 44(3), 47-58.

Litman, B.R. (1983). *Predicting Success of Theatrical Movies: An Empirical Study*. The Journal of Popular Culture, 16: 159-175.

Lindner, A.M., Lindquist, M. and Arnold, J. (2015). *Million Dollar Maybe? The Effect of Female Presence in Movies on Box Office Returns*. Sociol Inq, 85: 407-428.