THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF FINANCE

# Linear Regression, Mixture Modeling, and Gradient Boosting to Predict Box Office Revenue: Leveraging Machine Learning in Volatile Industries

JOSEPH JACOB PEVNER
SPRING 2023

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Finance
with honors in Finance

Reviewed and approved* by the following:

Lingzhou Xue
Associate Professor of Statistics
Thesis Supervisor

Brian Spangler Davis
Clinical Assistant Professor of Finance
Honors Adviser

* Electronic approvals are on file.

## ABSTRACT

Financial decision-making fundamentally relies upon our ability to accurately predict future cash flows, though in highly volatile markets, this poses an existential difficulty. This thesis explores the growing paradigm of applying regression and machine learning techniques to financial forecasting through a case-study of the notoriously erratic film industry. In this exploration, we pose three models of increasing complexity—a multiple linear regression, finite mixture model, and gradient boosting—to predict Domestic Box Office Revenue based upon several pre-release factors. Exploratory analysis, data wrangling, and feature engineering are employed upon a high-dimensional vendor-acquired dataset, emphasizing the importance of ensuring data quality prior to prediction. Each model is trained with five-fold cross-validation and five repetitions to promote robust and extrapolatable predictions. Comparing the evaluation metrics such as the Pearson Correlation Coefficient, Spearman's Correlation Coefficient, Mean Absolute Error, and Root Mean Squared Error across the three models demonstrates an increase in linearity and reduction in prediction error across an increase in model complexity. We find that the gradient boosted model is most effective in predicting revenues, approximately halving error from the baseline linear regression model, though the model poses difficulty in extracting general insights. We further submit finite mixture modeling as a balanced approach in maintaining algorithmic interpretability while generating accurate estimates. These findings demonstrate the ability of high-powered machine learning algorithms, such as expectation-maximization and gradient boosting, to forecast revenue in volatile financial environments.

# TABLE OF CONTENTS

# LIST OF FIGURES

## LIST OF TABLES

## ACKNOWLEDGEMENTS

I would like to thank the Schreyer Honors College for this incredible opportunity. I have been extremely lucky to grow and learn within this community for the past four years, and I will carry on the Schreyer mission and values for a lifetime. I am additionally immensely grateful for being awarded a Schreyer Honors College Research Grant which helped fund this study. Lauren Reiter, thank you for your kindness and assistance throughout the data acquisition process. You have gone above and beyond in supporting me and countless other Smeal students. Professor Brian Davis, thank you for your unwavering availability and kindness. You have been a wonderful teacher, mentor, and inspiration. Professor Lingzhou Xue, thank you for volunteering your time and expertise to this study. You have given me invaluable insights and guidance throughout this journey. To my family—Mom, Dad, Jessica, Nana, Papa, Mom-mom, Pop-pop, Uncle Chris, Uncle Pete, Aunt Tiffany, Aunt Marci, Evan, Aubrey, Uncle Frank, Uncle Barry, Yiera, Emily, and Justin—thank you for your infinite support and love.

# Chapter 1

## Introduction

The film industry is a unique market environment characterized by extreme uncertainty. Regardless of the amount of funding that studios may allocate in production, once a movie is released to the public, its success is entirely dependent on consumer behavior. Movies are an experiential good, one which does not fulfill an innate need, but rather is driven by hedonic value. Furthermore, repeat purchases (i.e., repeat theatrical viewings) are infrequent. A film's box office revenue is dependent upon a mass of consumers opting into purchasing a ticket, resulting from the perceived value of viewing (as created by a film's genre, marketing, etc.) exceeding cost (in 2023, an average of $10.45). Due to the experiential nature of movies, the popularity of a film, and in turn box office revenue, is subject to a contagion effect; the buzz surrounding a movie is a key driver in consumer choices. Furthermore, the inherent creation of a viewer in-group can lead to films having a quasi-network effect, as the perceived value of a film's viewership increases with the number of existent viewers.

Time and time again, big budget blockbusters become box-office flops and small budget films become breakout hits. While there certainly exists a relationship between budget and box office success, there is no such thing as a movie that is too big to fail. Historically, less than four out of ten movies have broken even from box office revenues. Studios must choose their films wisely—the creation of a movie requires an immense investment in capital, time, and personnel, and the financial success of a production studio is rooted in the success of its underlying films.

Our thesis is rooted in this paradigm, that studios must bridge the gap of uncertainty to produce the most profitable films. The intention of our analysis is to examine if a film's box office revenue, and in turn the financial success that it brings to a production studio, can be accurately predicted prior to its release. In examining this, we will create a series of statistical models to predict box office revenue. Comparing the efficacy of predictive models will yield valuable insights, not only in furthering the ability to predict the success of films, but also in better understanding the nature of data-driven financial predictions in highly volatile environments.

# Chapter 2

# Literature Review

## Economic Conditions

In order to succeed within the film industry, production companies must first solve the puzzle of consumer desire. Cooper-Martin (1991) demonstrated the importance of hedonic (pleasure) value, rather than utilitarian motives, in influencing a consumer's decision to purchase a movie ticket. Measuring the degree of importance attributed to various product-aspects by potential consumers faced with a slate of experiential goods found that movies garnered a significantly large consideration of hedonic value prior to purchase, more so than any other experiential good. Furthermore, when choosing between multiple films, consumers placed more weight upon subjective attributes (genre, tone, source material) than objective measures (theater location, ticket price, setting).

Walls and De Vany (2004) captured the economic conditions of film releases into the stable Paretian distribution $S(\alpha, \beta, \gamma, \sigma)$. Due to the experiential nature of movies, popularity—and in turn box office revenue—is highly subject to a contagion effect, as consumption-influencing information disseminates rapidly. The opening week of a film is especially indicative of its future success, since the movie's launch establishes the first, and often most influential, nodes in the viewer network. This environment generates a great amount of uncertainty regarding a film's success, exacerbated by a phenomenon described by Walls and De Vany as the "nobody knows" principle. This principle dictates that, due to the highly subjective and often chaotic nature of creative mediums, profitability within the film industry has infinite variance.

This infinite variance, dominated by extreme cases, leads to potentially huge disparities between a film's expected profit and modal profit which can severely mislead investment decisions.

## Revenue and Valuation

As per the Efficient Market Hypothesis, prices within the capital market reflect all available information. (Fama, 1970) This model, viewed in tandem with the "nobody knows" principal, yields complicated market implications for the equity valuation of production studios. Throughout the pre-production, production, and pre-release timeline of a film, pertinent information—both objective and subjective—is released into the public. The Efficient Market Hypothesis dictates that as a film comes closer to release and more information is publicized, the true value of the film becomes more and more accurately incorporated into the production studio's valuation. However, the infinite variance of profitability implicates that the vast majority of pertinent information comes with a films' release, after which the production studio would receive its greatest price adjustment.

The impact of pre-release and post-release news upon a film studio's valuation has been well documented. Einad and Ravid (2009) found a strong correlation between delays in film release dates and decreases in studio valuation. After a delay is announced, the degree to which the studio's equity valuation falls scales closely with the film's budget, but is uncorrelated to movie's eventual box office revenue, indicating that investors are more aware of cost-side risks of a film in the production phase.

Furthermore, Joshi and Hanssens (2009) examined and established the dual impact that pre-release marketing and opening weekend performance has upon studio valuation. These researchers identified advertisements as a key quality signal in building up investor expectations—that the more publicity a film receives, the higher the market will value the movie, and in turn the more capital will be invested in the production studio. Joshi and Hanssens found that post-release stock returns are a function of both the film's theatrical performance and the pre-release expectations of the film's performance.

These finding affirm the Efficient Market Hypothesis, as they suggest that the change in valuation following a film's release represents the market correcting a dissonance in a film's pre-release anticipated value and its post-release actualized value.

## Predictive Modeling

The first statistical model engineered to predict the success of film releases was created by Barry Litman (1983). Litman identified three essential areas which he deemed were deterministic in a film's success—creative decisions, scheduling/release timing, and marketing coverage. In quantifying these areas, creative decisions were measured by genre, MPAA rating, presence of superstars, production cost, and distributor; scheduling/release timing was measured by binary indicators for three peak release periods: November/December, March/April, and June/July/August; marketing coverage, however, was not accounted for in the model due to a lack of available data. The model also included two post-release metrics—critical reviews and Academy Award nominations/wins. After performing a multiple regression model and eliminating variables without statistical significance, Litman's model contained 7 variables of interest, 5 of which being indicator variables (horror/science fiction genre, major distributor, November/December release time, Academy Award nominee, and Academy Award winner) and

2 of which being quantitative (production cost and critic ratings). This model explained 48.5% of the variance within its 125 film dataset.

Litman's research was innovative in utilizing statistical models to predict the chaotic environment of the film industry and laid the groundwork for a plethora of subsequent studies. In 1996, Sawhney and Eliahsberg (1996) developed a stochastic model prioritizing parsimony to predict box theatrical visits. These researchers conceptually divided the total lifecycle of a consumer watching a film (time to adopt) into two metrics: the time to decide and the time to act. Within this behavioral framework, movie consumers encounter decision-influencing information which entices them to purchase a movie ticket (time to decide) and act upon this urge in purchasing a ticket (time to act) as two stochastic and independently occurring processes. Using three weeks of leading/simulated data, these two time-metrics were fitted into independent Gamma distributions, which were then cumulated into a single Binomial distribution in order to simulate the purchasing behavior of all potential moviegoers. The resultant model, BOXMOD-I, performed with an average prediction error of 11.23%; however, due to the three-week data requirements, this model is functionally best for informing post-release decisions.

In 2009, Yong Liu created a model which incorporated the word of mouth (WOM) surrounding a film, as measured by the volume and valence of posts the Yahoo Movies message board. Liu observed a "carryover" effect in WOM—the amount of buzz a film receives in a given week very strongly correlates the previous week. This observation affirms the contagion effect of consumption in the film industry. Testing the efficacy of predictive models before and after the inclusion of WOM found that the incorporation of a film's "buzz" added significant prediction power— reducing prediction error in opening week sales from 55% to 38% and in aggregate box office revenue from 61% to 47%.

Further studies have utilized online information sources such as user reviews (Chintagunnta et al., 2010), website promotion (Zufryden, 2000), and Wikipedia page activity (Marton et al., 2013) to build models based heavily upon the efficacy of a single predictor variable. Throughout these studies, the accuracy of box office revenue prediction is greatly improved by the incorporation of a proxy for consumer-interest. Though effective, these proxies are derived from post-distribution data which is not available in the production phase of financing.

This study will focus on expanding the groundwork laid by Litman, leveraging quantitative and qualitative attributes of films to predict their financial success, with more complex and robust statistical models. Implementing regression and machine learning techniques trained upon a wrangled and fully engineered dataset will not only enable the prediction of single film box office revenue, but also, examining and comparing the models in totality will provide insight upon latent behaviors within this industry of infinite variance.

# Chapter 3

# Data and Methodologies

## Bridging the Information Gap

One fundamental difficulty in utilizing analytics to model the film industry is a widescale lack of comprehensive, publicly available data. While some services, such as IMDb and Rotten Tomatoes, aim to democratize film information, these domains operate on a per-movie basis, providing access to information by user invocation rather than aggregating data. Manually transcribing from these sources is unsustainable at an analytically-viable scale. Additionally, the dynamic layout of these sites paired with a high degree of variation in available information across movies renders web-scraping inefficient, with a high likelihood of generating incomplete or inconsistent datasets.

To overcome this issue, we employed OpusData, an industry-leading data vendor which specializes in providing extensive and comprehensive film information. An Academic Extract on 02/08/2023 enabled us to access the data pertinent to our research on a large scale. Prior to data cleansing and feature engineering, this dataset contained 35,027 entries, each of which representing a unique film.

## Data Wrangling and Feature Engineering

Drawing from the current paradigms of film analysis and consumer behavior, we identified eight pertinent explanatory variables. These variables are: Production Budget, Running Time, Release Date, Opening Weekend Theaters, Sequel, Creative Type, Source, and Genre.

Including creative, technical, and commercial attributes determined across the preproduction, production, and pre-release phases of movie-making enabled our insights to be driven by the entirety of a film. We chose this approach to create a robust predictive model which accounted for each step of film creation. We identified our response variable of interest as Domestic Box Office Revenue.

After determining the attributes that would drive our predictions, we turned towards ensuring that the data was cleaned and prepared for modeling. While OpusData certainly provided a comprehensive dataset, there were still a number of errors to ameliorate prior to processing.

While initial exploration indicated that the dataset had almost no missing values, further investigation revealed that the dataset recorded missing values in a manner which R did not detect. For many numerical values (Production Budget, Opening Weekend Theaters, Running Time, and Domestic Box Office Revenue), missing values were stored as a value of 0. The use of these values as stored would result in erroneously left-skewed distributions. To correct this issue, missing values for explanatory variables (Production Budget, Opening Weekend Theaters, and Running Time), were inputted as the median of their respective distributions. For the response variable (Domestic Box Office Revenue), identified missing values were omitted. This process enabled us to maximally utilize the dataset without corrupting the validity of our variable of interest. Additionally, for categorical explanatory variables (Sequel, Genre, Creative Type, and Source), missing values were inputted as the text “NULL”. These values were removed from the dataset. Further data cleaning was conducted to improve readability, such as reformatting variable names and resequencing entries. After removing all incomplete entries, the dataset contained 4,325 unique films.

Furthermore, we engineered two new attributes to enhance the model—Release Month and Release Year. By decomposing the Release Date into a categorical month and quantitative year, we were able to incorporate the cyclical impact of release timing into our model while simultaneously tracking larger temporal trends of box office revenue. After encoding these new variables, the Release Date column was no longer necessary and was dropped from the dataset.

| Variable Name | Min | Q1 | Median | Q3 | Max | Arithmetic Mean | Standard Deviation |
|---|---|---|---|---|---|---|---|
| *Domestic Box Office* | *264* | *7,502,560* | *28,544,120* | *67,264,877* | *936,662,225* | *54,903,803* | *80,254,724* |
| Production Budget | 7,000 | 10,000,000 | 25,000,000 | 53,000,000 | 460,000,000 | 40,569,510 | 47,493,958 |
| Running Time | 41 | 95 | 106 | 119 | 220 | 109 | 19 |
| Opening Theaters | 1 | 21 | 2207 | 3018 | 4735 | 1839 | 1423 |
| Release Year | 1933 | 2003 | 2009 | 2015 | 2023 | 2008 | 10 |

**Table 1. Summary Statistics for Domestic Box Office and Quantitative Predictors**

Examining the quantitative variables was an essential step in ensuring data validity and better understanding the distribution of these factors. While Production Budgets and Domestic Box Office Revenue have similar first and second quartiles, Domestic Box Office Revenue demonstrates a strong right-skewed distribution. Furthermore, the sheer variability of Domestic Box Office Revenue is demonstrated by its standard deviation of over $80 million. Half of the movies in the dataset have running times between just over an hour and a half and two hours. Additionally, the dataset contains ninety years of movies, from 1933 to 2023.

Exploratory data analysis further revealed that the variables Source and Creative types had a large number of unique levels. In order to ensure that there was sufficient data within each level of these factors, we consolidated levels which applied to less than twenty unique films into the single level “Other”. For Source, this combined Ballet, Musical Group, Movie, Musical/Opera, Religious Text, Song, Theme Park Ride, Toy, Web Series, and Compilation, into Other with a total of 80 entries. For Creative Type, this combined Concert Performance and Multiple Genres into Other with a total of 15 entries. The consolidation aided in reducing the dimensionality of the dataset, decreasing the likelihood of overfitting, and improving computational efficiency.

The final step of featuring engineering prior to model-building was the creation of dummy variables. This process involves converting categorical variables into binary indicators that take on values of 0 or 1 to represent the absence or presence of particular attributes. For instance, the Genre variable had twelve unique values, including Action, Comedy, and Musical. In the creation of dummy variables, this singular categorical attribute of Genre was converted into twelve dummy variables, each corresponding to a different unique genre such that any given movie in the dataset has a value of 1 in one of the twelve Genre dummy variables, indicating that the movie has that genre, and a value of 0 in the other eleven Genre dummy variables. After completing this process, the dataset's four categorical variables inhabited forty-seven dummy variables.

| Variable Name | Levels |
|---|---|
| Source | Original Screenplay, Comic/Graphic Novel, Factual Book/Article, Fiction Book/Short Story, Folk Tale/Legend/Fairytale, Game, Musical Play, Real Life Events, TV, Remake, Spinoff, Other |
| Creative Type | Contemporary Fiction, Dramatization, Factual, Fantasy, Historical Fiction, Kids' Fiction, Multiple, Science Fiction, Superhero |
| Genre | Drama, Action, Adventure, Black Comedy, Comedy, Documentary, Horror, Musical, Romantic Comedy, Thriller/Suspense, Western, Other |
| Release Month | January, February, March, April, May, June, July, August, September, October, November, December |
| Sequel | Yes, No |

**Table 2. Categorical Predictors of Interest**

## Model Selection

We identified and implemented three models of increasing complexity to explore different approaches to analyzing the highly erratic environment of box office revenue. The first model is a multiple linear regression. The second model is a finite mixture model. The third model is gradient boosting.

## Multiple Linear Regression

Linear regression is a technique used to model the relationship between a numerical response variable and one or more explanatory variables by forming a linear equation. One concern in creating a multiple linear regression model is multicollinearity– a condition wherein several quantitative predictors are highly correlated. Including collinear variables in a regression model is bad practice, as it leads to equations that have redundant variables with potentially misleading coefficients.

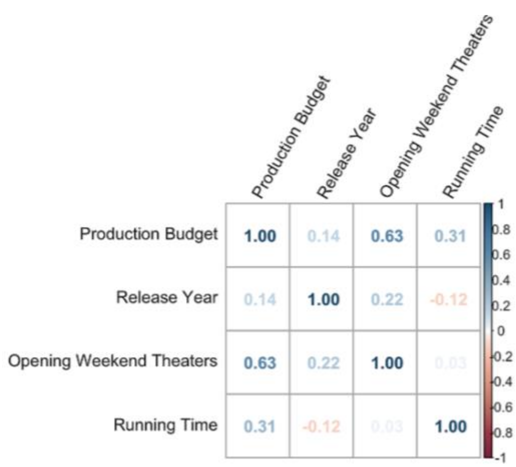


**Figure 1. Correlation Matrix of Quantitative Predictors**

Conventionally, a Pearson correlation coefficient ($r$) with an absolute value ($|r|$) between the values of 0.7 and 1 is considered indicative of a strong linear relationship. As evidenced in the correlation matrix, no two of our quantitative predictors have Pearson correlation coefficients which fall within this range, suggesting that the condition of noncollinearity is fulfilled. The largest $r$ value identified within the matrix is between Production Budget and Opening Weekend Theaters, with a value of 0.63. While these two variables are certainly related—movies with higher production budgets tend to have higher distribution budgets— they ultimately capture the impact of two fundamentally different stages of a film's lifecycle and neither conceptually nor statistically contribute to modelling redundancy.

Multicollinearity is also a concern in the implementation of categorical variables. In a phenomenon called the Dummy Variable Trap, the inclusion of all dummy variables of a categorical attribute leads to perfect correlation ($r = 1$), and in turn multicollinearity, across the given dummy variables. To circumvent this, we designated reference levels for each categorical variable to be excluded from the linear regression. The presence of a reference level is indicated in the model by all dummy variables for a given category having a value of 0. In interpreting the linear regression model, the coefficients for dummy variables represent their relative impact upon Domestic Box Office Revenue vis-à-vis the reference level.

| **Categorical Variable** | **Reference Level** |
|---|---|
| Source | Original Screenplay |
| Creative Type | Contemporary Fiction |
| Genre | Drama |
| Release Month | January |
| Sequel | No |

**Table 3. Reference Level Assignment for Categorical Variables**

In creating the linear model, we utilized Repeated k-Fold Cross Validation via the caret package (with $k = 5$ and $repeats = 5$). This process splits the data into five sets of equal size and procedurally cycles through each fold, using 80% of the data to train and 20% to test. Performing k-fold cross validation ensures that the resultant linear model is robust and helps prevent overfitting. Repeating this procedure five times further reduces the risk of obtaining a biased estimate due to chance.

**Finite Mixture Model**

Rather than fitting a singular predictive line onto the dataset, finite mixture modeling supposes that there are multiple, latent subpopulations within our data which can be independently fit with linear equations. Finite mixture models are often effective in environments with complex heterogenous data that does not conform to a single clear distribution.

In fitting mixture models to the film dataset, we utilized the FlexMix package in R. FlexMix implements the Expectation-Maximization (EM) algorithm, which iteratively estimates the parameters of a mixture model by alternating between E-steps and M-steps. In the E-step, FlexMix computes the posterior probabilities of each observation belonging to each component of the mixture model given the parameter estimates. In the M-step, FlexMix updates the parameter estimates by maximizing log-likelihood given the posterior probabilities.

We passed a General Linear Model (GLM) to FlexMix such that Domestic Box Office was framed as a function of all identified explanatory variables. Furthermore, we added a control upon the E-step by setting the minimum prior probability for components to 0.15. This value was tuned to maximize the components used in any given step while preventing errors due to numerical instability. We retained the designated dummy variable reference levels from the multiple linear regression. Additionally, we employed Repeated k-Fold Validation with five folds and five repetitions in FlexMix to promote an extrapolatable and robust model.

Since the number of latent subpopulations in the film dataset is unknown, we iterated FlexMix across starting component values ($k_0$) 1 to 5. After fitting five unique mixed models, we selected the model of best fit based upon the highest Integrated Completed Likelihood (ICL). ICL is a criterion used for evaluating and comparing mixture models which balances fit against complexity. The $k_0$ of the lowest ICL model provides insight into the most effective number of components for modeling the dataset.

**Gradient Boosting**

Gradient Boosting is a machine learning technique which iteratively forms a series of weak models, each correcting the error of the previous model. Specifically, Gradient Boosting Machines (GBMs) utilize gradient descent to minimize a designated loss function. Due to its adaptability and efficiency, Gradient Boosting works well in modeling high-dimensional, noisy data.

In utilizing a GBM to predict Domestic Box Office Revenue, we used the XGBoost package in R. XGBoost is a popular implementation of Gradient Boosting which uses advanced regularization within a Gradient Boosting framework to increase efficiency and produce more generalizable models. For the training of our model, we designated the loss function as Root Mean Square Error (RMSE). Using RMSE as opposed to Mean Absolute Error (MAE) leads to a predictive model which more strongly penalizes large regressions, which is advantageous for the outlier-driven nature of the film industry. We tuned hyperparameters using a grid search approach, resulting in the following:

| Hyperparameter | Value |
|---|---|
| Gamma | 0 |
| ETA | 0.3 |
| Maximum Depth | 5 |
| Minimum Child Weight | 1 |
| Subsample | 0.8 |
| Column Sample by Tree | 0.8 |

**Table 4. Tuned XGBoost Hyperparameter Values**

To train the model, we performed a k-Fold Cross Validation with five folds and a maximum of 10,000 rounds. We additionally implemented a control to end modeling if the training cycle generated 25 successive rounds with no improvement to prevent overfitting and needless computation. Once the training had completed, we extracted the best iteration based upon the minimum RMSE.

# Chapter 4

# Findings

## Multiple Linear Regression

The multiple linear regression yielded a singular linear equation for predicting box office revenue. The output of this model are as follows:

| Variable | Coefficient Estimate | Standard Error | P-Value |
|---|---|---|---|
| Intercept | 1600484072 | 182357952 | < 2e-16 |
| Production Budget | 0.7763278 | 0.029 | < 2e-16 |
| Running Time | 487653 | 54655 | < 2e-16 |
| Opening Weekend Theaters | 11248 | 867 | < 2e-16 |
| Release Year | -826281 | 90266 | < 2e-16 |
| Release Month – February | 3738242 | 4600193 | 0.416 |
| Release Month – March | 5182898 | 4467220 | 0.246 |
| Release Month – April | 3482613 | 4537085 | 0.443 |
| Release Month – May | 19990367 | 4759792 | 2.73e-5 |
| Release Month – June | 19585409 | 4521151 | 1.51e-05 |
| Release Month – July | 13052540 | 4526739 | 0.004 |
| Release Month – August | 876410.265 | 4408011 | 0.842 |
| Release Month – September | 998992 | 4436838 | 0.822 |
| Release Month – October | 737180 | 4284972 | 0.863 |
| Release Month – November | 10007988 | 4422595 | 0.0237 |
| Release Month – December | 18062859 | 4389320 | 3.94e-05 |
| Creative Type – Dramatization | -3025682 | 5187409 | 0.560 |
| Creative Type – Factual | 29183725 | 29329725 | 0.320 |
| Creative Type – Fantasy | -2793404 | 3783882 | 0.460 |
| Creative Type – Historical Fiction | -8481213 | 3122516 | 0.007 |
| Creative Type – Kids Fiction | 11366000 | 4660374 | 0.015 |
| Creative Type – Multiple | -35749229 | 25317042 | 0.158 |
| Creative Type – Science Fiction | 2946092 | 3230349 | 0.362 |
| Creative Type – Superhero | 61735376 | 7689934 | 1.27e-15 |
| Source – Comic/Graphic Novel | -3750821 | 6117846 | 0.540 |
| Source – Factual Book/Article | 4846172 | 5946823 | 0.415 |
| Source – Fiction Book/Short Story | -5893204 | 2380226 | 0.013 |
| Source – Folk Tale/Legend/Fairytale | -7749431.8 | 8989855 | 0.389 |
| Source – Game | -32791144 | 9060223 | 3.00e-4 |
| Source – Play | 3058548 | 7060248 | 0.665 |

| | | | |
|---|---|---|---|
| Source – Real Life Events | -5291035 | 5488979 | 0.335 |
| Source – Short Film | -7370998 | 11491081 | 0.521 |
| Source – TV | -10450373 | 4999414 | 0.037 |
| Source – Other | -10527471 | 6617083 | 0.112 |
| Source – Remake | -4647924 | 4296919 | 0.279 |
| Source – Spin Off | 19061811 | 10392586 | 0.067 |
| Genre – Action | -19117682 | 3393520 | 1.88e-08 |
| Genre – Adventure | 3890398 | 4125260 | 0.346 |
| Genre – Black Comedy | -3803013 | 6446550 | 0.555 |
| Genre – Comedy | 2816874 | 2996608 | 0.347 |
| Genre – Documentary | -14509984 | 29334891 | 0.621 |
| Genre – Horror | 2123030 | 3792348 | 0.576 |
| Genre – Musical | 16885366 | 7358186 | 0.022 |
| Genre – Other | -10827811 | 29239324 | 0.711 |
| Genre – Romantic Comedy | 476774 | 4416021 | 0.914 |
| Genre – Thriller/Suspense | -7393388 | 3230786 | 0.022 |
| Genre – Western | -7982001 | 9127878 | 0.382 |
| Sequel – Yes | 24647253 | 2843657 | < 2e-16 |

**Table 5. Multiple Linear Regression Output**

The value of the Intercept, 1,600,484,072, theoretically represents the estimated mean Domestic Box Office Revenue for a film with values of zero for all predictor variables, though in practice, this would never be the case. Rather, the Intercept is simply a corrective term within the model to add once all coefficients have been applied.

The Production Budget coefficient, 0.776, suggests that, holding all other variables constant, every additional $1 spent on production translates to around $0.78 in Domestic Box Office Revenue. This coefficient affirms the paradigm within which this thesis lies—that simply increasing the budget of a film is not enough to break even, rather, studios must carefully and intentionally control for other aspects of the production and release cycle.

The Opening Weekend Theater coefficient implies that every additional theater that a movie screens within during the first weekend results in an increase of $11,248 in Domestic Box Office Revenue. This finding falls in line with Walls and De Vany's identification of a highly deterministic contagion effect within the film industry which originates from opening week.

The Release Year coefficient suggests that Domestic Box Office Revenue decreases around $826,281 per year, a phenomenon potentially attributable to the proliferation of alternative forms of film consumption, namely streaming services.

In interpreting the categorical variable coefficients, it is important to keep in mind that the values represent the impact on Domestic Box Office Revenue relative to the reference levels.

The linear regression yields a model with a coefficient of determination ($R^2$) value of 0.544. This suggests that just over half of the variance in Domestic Box Office Revenue, 54.4%, is explained by a linear model with the identified predictors. After correcting for the number of predictors in the model, the adjusted coefficient of determination ($Adjusted\ R^2$) has a value of 0.5379, only slightly lower than $R^2$, suggesting that the model is not overspecified.

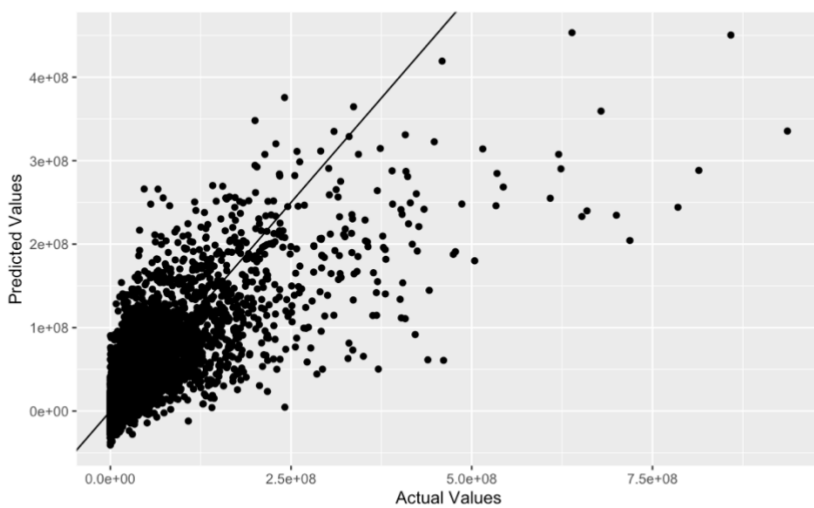


**Figure 2. Prediction Error Plot for Multiple Linear Regression**

The Prediction Error Plot for this model shows a positive relationship between the predicted and actual values of Domestic Box Office Revenue. However, a wide spread of points indicates that there is a substantial amount of variance left unexplained by the linear regression.

**Finite Mixture Model**

After iterating through five potential mixture models, the model which yielded the lowest Integrated Completed Likelihood, 163,622, was that which had the starting component value, $k_0 = 2$. This $k_0$ value indicates that two clusters, and in turn two corresponding linear models, are sufficient in capturing the underlying patterns of Domestic Box Office Revenue. Within this two-component mixture model, the first cluster, Component 1, encapsulated 71% of the dataset (3301 films), and the second cluster, Component 2, encapsulated 29% of the dataset (1024 films). The FlexMix-estimated coefficients of these two components are as follows:

| **Component 1** | | **Component 2** | |
|---|---|---|---|
| **Variable** | **Coefficient Estimate** | **Variable** | **Coefficient Estimate** |
| Intercept | -900794862.5 | Intercept | 762648038.4 |
| Production Budget | 0 | Production Budget | 0 |
| Running Time | 1651484 | Running Time | 182435.6 |
| Opening Weekend Theaters | 0 | Opening Weekend Theaters | 11502.68 |
| Release Year | 377709.1 | Release Year | -388245 |
| Release Month – February | 31369569 | Release Month – February | 1112640 |
| Release Month – March | 33247836 | Release Month – March | 1175410 |
| Release Month – April | 38141390 | Release Month – April | 0 |
| Release Month – May | 58522642 | Release Month – May | 2767081 |
| Release Month – June | 58180084 | Release Month – June | 5152786 |
| Release Month – July | 47191700 | Release Month – July | 3872914 |
| Release Month – August | 23248591 | Release Month – August | -622148 |
| Release Month – September | 10143884 | Release Month – September | -750872 |
| Release Month – October | 9224277 | Release Month – October | -1256532 |

| | | | |
|---|---|---|---|
| Release Month – November | 37511180 | Release Month – November | 2196763 |
| Release Month – December | 55819457 | Release Month – December | 7757376 |
| Creative Type – Dramatization | 0 | Creative Type – Dramatization | 0 |
| Creative Type – Factual | 69301529.09 | Creative Type – Factual | -5502945.5 |
| Creative Type – Fantasy | 22686121.08 | Creative Type – Fantasy | -1975409.75 |
| Creative Type – Historical Fiction | -15746074.21 | Creative Type – Historical Fiction | -1420900.27 |
| Creative Type – Kids Fiction | 55539116.95 | Creative Type – Kids Fiction | -3728429.41 |
| Creative Type – Multiple | -39346703.31 | Creative Type – Multiple | -16223050.5 |
| Creative Type – Science Fiction | 46271188.83 | Creative Type – Science Fiction | -1352322.16 |
| Creative Type – Superhero | 110771453.9 | Creative Type – Superhero | 1549848.63 |
| Source – Comic/Graphic Novel | 0 | Source – Comic/Graphic Novel | 711552.386 |
| Source – Factual Book/Article | 865734.164 | Source – Factual Book/Article | 2695169.2 |
| Source – Fiction Book/Short Story | -11961577.66 | Source – Fiction Book/Short Story | 1196584.04 |
| Source – Folk Tale/Legend/Fairytale | 29031795.26 | Source – Folk Tale/Legend/Fairytale | 4215729.57 |
| Source – Game | -51308274.49 | Source – Game | -209010.484 |
| Source – Play | -1755814.889 | Source – Play | 6183328.33 |
| Source – Real Life Events | -26303756.54 | Source – Real Life Events | 851153.726 |
| Source – Short Film | -52699257.03 | Source – Short Film | 7943126.58 |
| Source – TV | -26488874.65 | Source – TV | 4132970.78 |
| Source – Other | -5599359.441 | Source – Other | -3204511.1 |
| Source – Remake | 2119810.701 | Source – Remake | 2082165.01 |
| Source – Spin Off | 41488184.01 | Source – Spin Off | 18101339.1 |
| Genre – Action | 17445770.26 | Genre – Action | -1686682.46 |
| Genre – Adventure | 54955679.05 | Genre – Adventure | 5701052.63 |
| Genre – Black Comedy | -27566334.92 | Genre – Black Comedy | 1938796.85 |
| Genre – Comedy | 28719044.8 | Genre – Comedy | -107272.969 |
| Genre – Documentary | -56270129.47 | Genre – Documentary | 7725668.16 |
| Genre – Horror | 13361898.5 | Genre – Horror | -144356.01 |

| | | | |
|---|---|---|---|
| Genre – Musical | 39960642.85 | Genre – Musical | -680758.854 |
| Genre – Other | -30189421.51 | Genre – Other | 8206497.48 |
| Genre – Romantic Comedy | 12166686.77 | Genre – Romantic Comedy | 594369.612 |
| Genre – Thriller/Suspense | 11083125.01 | Genre – Thriller/Suspense | -788189.377 |
| Genre – Western | 0 | Genre – Western | 881292.907 |
| Sequel – Yes | 54997156 | Sequel – Yes | 8656515 |

**Table 6. Finite Mixture Model Output**

The Intercepts of these two components demonstrate a considerable disparity, Component 1's is -900,794,862.5 and Component 2's is 762,648,038.4, which is indicative of fundamental differences in behavior between the underlying subpopulations captured within the mixture model. This divergence highlights the necessity of separate modeling of the two components. Furthermore, it is important to remain mindful of this discrepancy while comparing coefficients across the models.

It should be noted that very low negative intercepts can inflate the values of dummy coefficients, which increases the risk for erroneous interpretations. For instance, in Component 1, the coefficient for Release Month – December is equal to 55,819,457. This value cannot be interpreted as evidence that the mere act of releasing a film in December increases its Domestic Box Office Revenue by over $55 million. Rather, this value is relative to and dependent upon all other variables in the component model.

Notably, the coefficient for Production Budget across both components of the mixture model is equal to zero, suggesting that a film's budget is not a significant predictor of box office revenue. This is a fundamental departure from the linear regression model wherein revenue was identified as a highly significant predictor of Domestic Box Office Revenue. A mixture model may be advantageous in situations wherein the Production Budget of a film is unknown.

Additionally, for Component 1, Opening Weekend Theaters has a coefficient of zero, whereas Component 2 has a coefficient similar to that generated in the multiple linear regression. Viewing this in conjunction with a lack of coefficient for Production Budget, Component 1 may be a model that can functionally predict Box Office Revenue prior to the distribution-phase of a film.

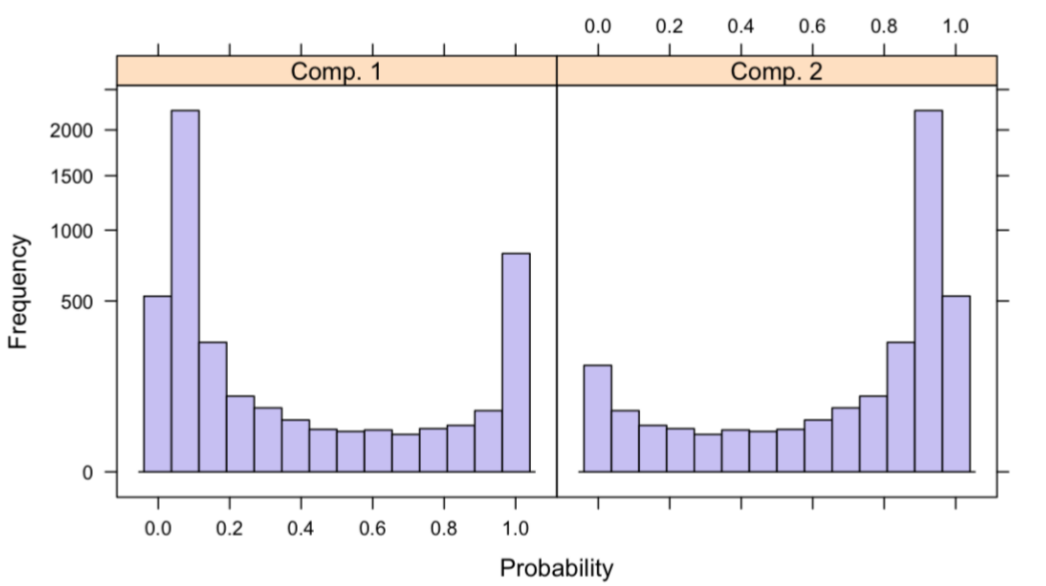


**Figure 3. Rootogram of Posterior Probabilities > 1e-04**

Rootograms provide valuable insight in visualizing how effectively a mixture model separates components. A peak near 1.0 on a component rootogram indicates that the component is well separated from the others. The rootogram for Component 2 shows a prominent peak between 0.8 and 1.0, indicating that it is well-separated from other components. However, the rootogram for Component 1 has a large peak between 0.0 and 0.2, which suggests substantial overlap. This indicates that, even after accounting for k-Fold Cross Validation with repeats and ICL-driven model selection, a mixture model may struggle to generate fully distinct components for high-dimensional datasets.

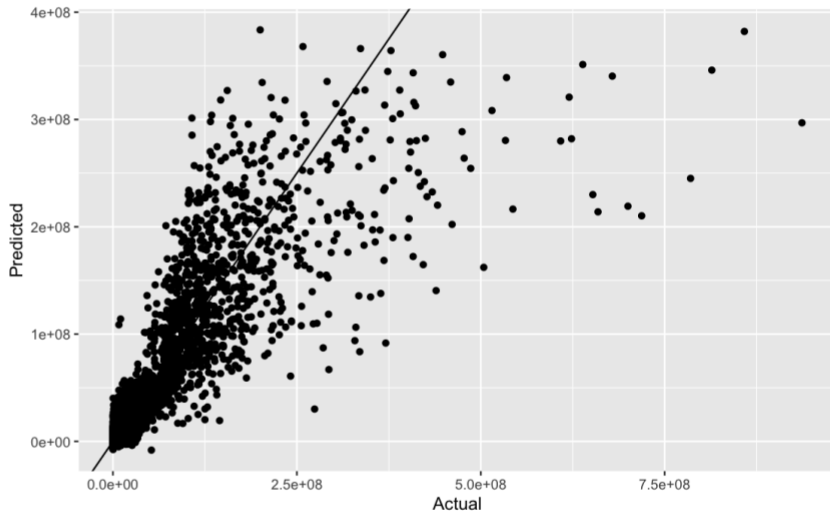


**Figure 4. Prediction Error Plot for Finite Mixture Model**

The Prediction Error Plot for the mixture model shows a clear improvement over the linear regression. Points are distributed more evenly around the best-fit line, indicating that a two-component mixture model improves our ability to explain variance within the film industry. However, there is a fanning effect as Actual Domestic Box Office Revenue increases, indicating that the model is less effective at predicting extreme values.

## Gradient Boosting

After 104 iterations, XGBoost converged onto a model which most efficiently reduced the loss function, RMSE, while accounting for the limit on overfitting. This yielded a Root Mean Squared Error of 29,296,076. The Gradient Boosted model can be visualized with the following tree diagram:

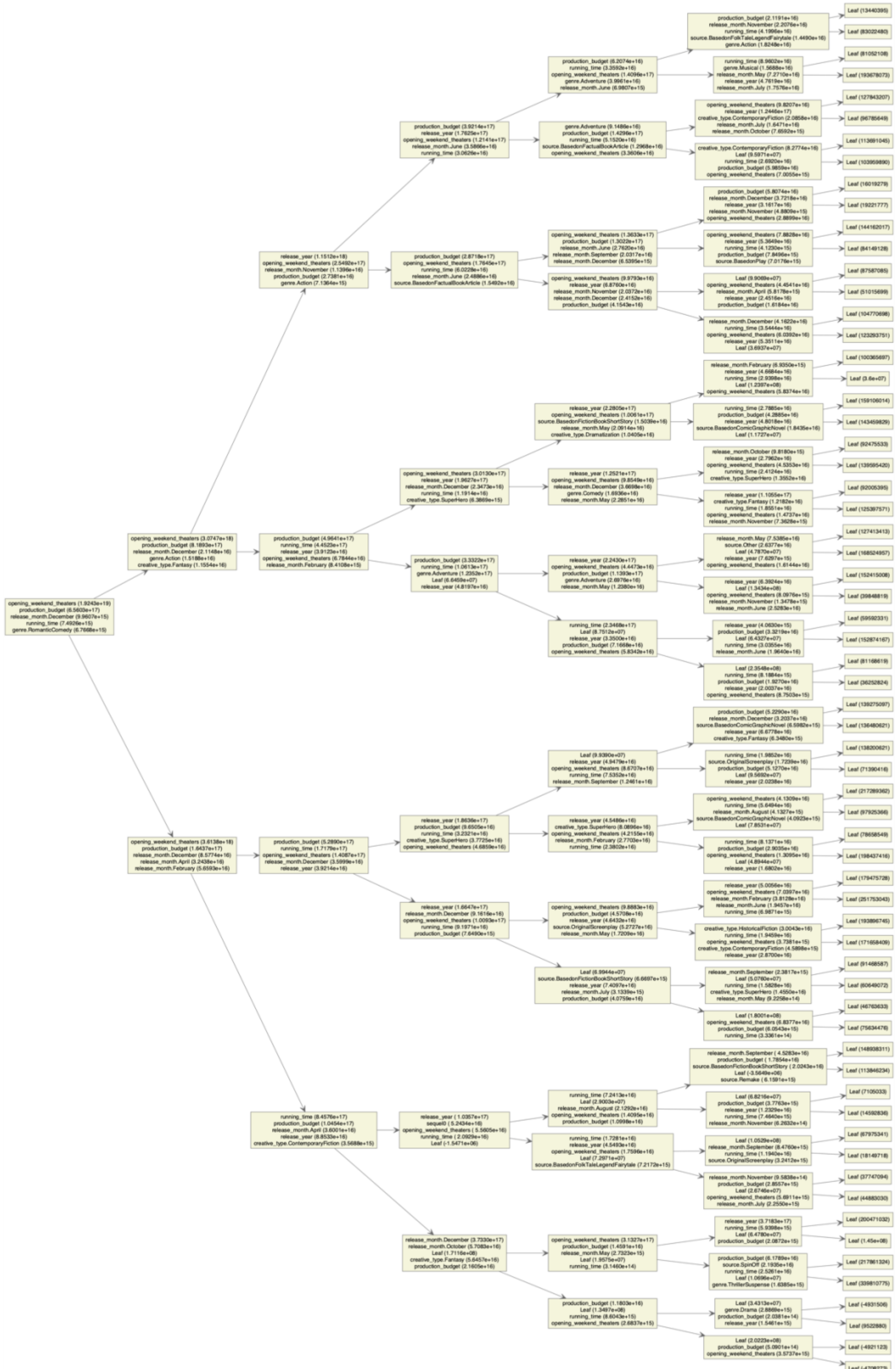

**Figure 5. Gradient Boosted Model Output**

The XGBoost model does not offer as straightforward of an interpretation as the previous linear models, though the tree diagram gives insight into how the gradient boosted model generates a prediction. For a given film, the model uses the predictor variables to traverse the decision tree at each split until it reaches a leaf node at which point the values of traversed nodes are summed to calculate a final prediction.

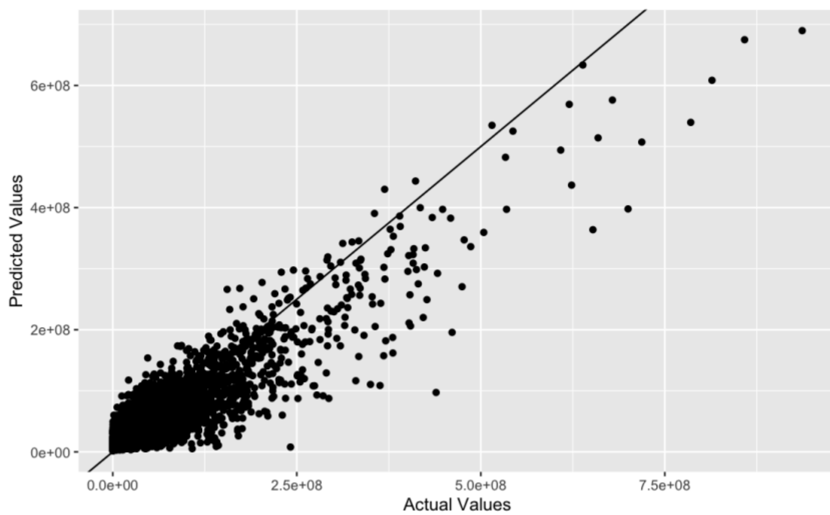


**Figure 6. Prediction Error Plot for Gradient Boosted Model**

The Prediction Error Plot for XGBoost shows a further improvement over the previous models in reducing the error of predictions across the dataset. In particular, the model shows a significantly lower error for high Domestic Box Office Revenue values, indicating that the model is better at predicting very successful films. This improvement can be attributed to the designation of Root Mean Squared Error as the loss function, which strongly penalized large regressions. It is worth noting that for these extreme values, the model tends to underestimate Domestic Box Office Revenue, which may be advantageous in situations that demand conservative financial estimation.

**Model Comparison**

Viewing evaluation metrics allow us to directly compare the ability to predict Domestic Box Office Revenue across the three models.

| Model | Pearson Correlation Coefficient ($r$) | Spearman's Correlation Coefficient ($\rho$) | Mean Absolute Error | Root Mean Squared Error |
|---|---|---|---|---|
| Linear Regression | 0.738 | 0.785 | 32,400,479 | 54,186,539 |
| Mixture Model | 0.839 | 0.901 | 21,398,113 | 43,797,083 |
| Gradient Booted | 0.933 | 0.875 | 18,715,434 | 29,296,076 |

**Table 7. Evaluation Metrics Across Models**

For all three models, both the Pearson and Spearman's Correlation Coefficients fall above 0.7, suggesting strong relationships between the predicted and actual values.

The Pearson Correlation Coefficient increases steadily from Linear Regression to Gradient Boosted, indicating that an increase in model complexity leads to an increase in the linearity between predicted and actual values of Domestic Box Office Revenue.

Spearman's Correlation Coefficient, conversely, is highest for the Mixture Model which indicates that this model performs best in predicting the ordinality of Domestic Box Office Revenue. A two-component mixture model may be preferable in situations where the relative performance of films is more important than the revenue of an individual movie.

Both Mean Absolute Error and Root Mean Squared Error exhibit significant reductions across the three models with minimum errors in the Gradient Boosted model. Viewing the marginal error reduction across complexity, the movement from Linear Regression to a Mixture Model leads to the largest reduction in MAE—over $10 million—whereas the added complexity of the Gradient Boosted model only reduces MAE by an additional $3 million. Conversely, RMSE maintains large reductions—over $10 million—across all three models.

# Chapter 5

## Conclusion

Through data wrangling, feature engineering, and repeated statistical modeling, we have generated three models that predict Domestic Box Office Revenue with high linearity based upon the pre-release factors of Production Budget, Running Time, Opening Weekend Theaters, Release Year, Release Month, Sequel, Source, Creative Type, and Genre. These models–multiple linear regression, finite mixture modeling, and gradient boosting–demonstrate large reductions in prediction error, as indicated by both MAE and RMSE, across increases in model complexity.

The gradient boosted model has the highest accuracy in capturing the theoretically infinite financial variance of the film industry, demonstrating the effectiveness of supervised machine learning techniques in predicting erratic environments. However, due to its intricacy, it is difficult to extract specific insights from this model. Conversely, the Litman-derived multiple linear regression provides a straightforward approach that clearly demonstrates the effect of each factor upon Domestic Box Office Revenue, though this model only explains around half of the variance in the dataset. The finite mixture model offers a balanced approach, utilizing Expectation-Maximization to generate an interpretable model with high error reduction. In better understanding the implications of the mixture model, further research is necessary to explore the two identified sub-populations and improve overall component generation.

These findings contribute to the growing paradigm integrating machine learning techniques with financial planning and decision making. Leveraging high-powered algorithms such as Expectation-Maximization and Gradient Boosting can reveal hidden relationships within erratic economic environments, challenging the longstanding convention that “nobody knows”.

**BIBLIOGRAPHY**


Barnes, B. (2022, July 1). Pandemic-battered movie theaters are feeling good after a strong June. *The New York Times*. https://www.nytimes.com/live/2022/07/01/business/economy-news-inflation-stocks

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing*, *67*(4), 103–117.

Bertoletti, M., Friel, N., & Rastelli, R. (2015). *Choosing the number of clusters in a finite mixture model using an exact Integrated Completed Likelihood criterion* (arXiv:1411.4257). arXiv. https://doi.org/10.48550/arXiv.1411.4257

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions On*, *22*, 719–725. https://doi.org/10.1109/34.865189

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., & implementation), Xgb. contributors (base Xgb. (2023). *xgboost: Extreme Gradient Boosting* (1.7.3.1). https://CRAN.R-project.org/package=xgboost

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets. *Marketing Science*, *29*(5), 944–957. https://doi.org/10.1287/mksc.1100.0572

Cooper-Martin, E. (1991). Consumers and Movies: Some Findings on Experiential Products. *ACR North American Advances*, *NA-18*. https://www.acrwebsite.org/volumes/7187/volumes/v18/NA-18/full

De Vany, A., & Walls, W. D. (1999). Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office? *Journal of Cultural Economics*, *23*(4), 285–318. https://doi.org/10.1023/A:1007608125988

Einav, L., & Ravid, S. A. (2009). Stock market response to changes in movies' opening dates. *Journal of Cultural Economics*, *33*, 311–319. https://doi.org/10.1007/s10824-009-9105-3

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

Gruen, B., Leisch, F., Sarkar, D., Mortier, F., & Picard, N. (2023). *flexmix: Flexible Mixture Modeling* (2.3-19). https://CRAN.R-project.org/package=flexmix

Hennig-Thurau, T., Houston, M. B., & Walsh, G. (2007). Determinants of motion picture box office and profitability: An interrelationship approach. *Review of Managerial Science*, *1*(1), 65–92. https://doi.org/10.1007/s11846-007-0003-9

Joshi, A., & Hanssens, D. (2009). Movie Advertising and the Stock Market Valuation of Studios: A Case of "Great Expectations?" *Marketing Science*, *28*, 239–250. https://doi.org/10.1287/mksc.1080.0392

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

Litman, B. R. (1983). Predicting Success of Theatrical Movies: An Empirical Study. *The Journal of Popular Culture*, *16*(4), 159–175. https://doi.org/10.1111/j.0022-3840.1983.1604_159.x

Liu, Y. (2006). Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. *Journal of Marketing*, *70*(3), 74–89. https://doi.org/10.1509/jmkg.70.3.074

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLOS ONE*, *8*(8), e71226. https://doi.org/10.1371/journal.pone.0071226

*OpusData*. (n.d.). Retrieved April 2, 2023, from https://www.opusdata.com/home.php

Sawhney, M. S., & Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science*, *15*(2), 113–131.

Terry, N., Butler, M., & De'Armond, D. (2011). *THE DETERMINANTS OF DOMESTIC BOX OFFICE PERFORMANCE IN THE MOTION PICTURE INDUSTRY*. 12.

*The Numbers—Movie Market Summary 1995 to 2023*. (n.d.). The Numbers. Retrieved March 17, 2023, from https://www.the-numbers.com/market/

Walls, W., & De Vany, A. (2004). Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. *Journal of Economic Dynamics and Control*, *28*, 1035–1057. https://doi.org/10.1016/S0165-1889(03)00065-4

Zufryden, F. (2000). New Film Website Promotion and Box Office Performance. *Journal of Advertising Research*, *40*(1–2), 55–64. https://doi.org/10.2501/JAR-40-1-2-55-64

## ACADEMIC VITA

# Joseph Pevner

## EDUCATION

**The Pennsylvania State University | Smeal College of Business | Schreyer Honors College** **University Park, PA**
B.S. in Finance, Minors in Statistics & Information Systems Management **Class of 2023**
Thesis: *Linear Regression, Mixture Modeling, and Gradient Boosting to Predict Box Office Revenue*

## PROFESSIONAL EXPERIENCE

**PricewaterhouseCoopers** **Philadelphia, PA**
*Intern — Transfer Pricing Consulting* *June 2022 – August 2022*

- Employed proprietary software in analyzing financial statements and web scraping business descriptions, appending multi-million-entry databases to find best-fit companies most comparable to a tested party
- Utilized Excel and VBA to automate spreadsheet population, increasing efficiency 5x on client assignment
- Prepared industry analyses, transmittal letters, transfer pricing reports, planning reports, and other documentation

**Chubb Ltd.** **Philadelphia, PA**
*Intern — Enterprise Risk Management* *May 2021 – May 2022*

- Shortened data entry timeline for company dashboard 1000x by automating the process of collecting, formatting, and uploading information from third-party data vendors via SQL queries and Excel functions
- Improved risk modeling of unreported client-supplier relationships by aggregating and querying bill of lading data
- Built and presented supply chain network diagrams by integrating Excel analysis with Power BI

**Summit Risk Management & Insurance** **Horsham, PA**
*Intern — Data Analysis* *May 2020 – August 2020*

- Created dashboard to inform adjuster decisions regarding client-defendant pairing, populated via R analysis
- Designed Excel spreadsheet to automate employee wage adjustments, aggregate performance reviews

***The Daily Collegian*** **University Park, PA**
*Data Analyst* *January 2020 – Present*

- Analyze activity of 10,000+ daily users of student-run news site via Google Data Studio and Tableau
- Proposed and lead multimedia project resulting in sustained 25% increase in pageviews for 18-24 demographic

## LEADERSHIP ROLES

**Four Diamonds THON** **University Park, PA**
*Fundraising Development Coordinator* *May 2022 – Present*

- Establish and enforce fundraising, advertising, and documentation guidelines for the largest student-run philanthropy in the world benefiting pediatric cancer research and care with annual proceeds of over $10 million
- Perform ad hoc analyses for 200+ student organizations, design descriptive and predictive graphics with Tableau
- Spearheaded user interface redesign, formed relationships with local businesses, and revamped process of reviewing and approving fundraisers, resulting in a 2x increase from last year in funds raised to date (+$150,000)

*Fundraising Development Captain* *September 2021 – May 2022*

- Responsible for interviewing and selecting committee of 25 volunteers, leading weekly meetings

**Philosophy/Psychology/Sociology 120N: Knowing Right from Wrong** **University Park, PA**
*Undergraduate Teaching Assistant* *August 2020 – May 2021*

- Maintained weekly grading of assignments and provided instructional aid for 200+ students
- Facilitated upscaling of curriculum for a tenfold increase in class size, trained and supervised 5 new TAs

## HONORS AND AWARDS

The Evan Pugh Senior Scholar Award, The Evan Pugh Junior Scholar Award, The President Sparks Award,
The President's Freshman Award, Dean's List 7/7 Semesters, Schreyer Academic Excellence Scholarship

## SKILLS AND INTERESTS

- **Programming Languages/Software:** SQL, R, SAS, Python, Java, RapidMiner, VBA, Excel, Power BI, Tableau
- **Statistics:** data collection/wrangling/visualization/integration/privacy, database management, clustering, classification, regression/correlation/comparison tests, supervised/unsupervised machine learning
- **Finance:** portfolio optimization, @Risk, volatility estimation, GARCH, EWMA, option pricing models, Monte Carlo simulations, forecasting financial statements/profits/cashflows, capital budgeting, Black-Scholes modeling