THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF TELECOMMUNICATIONS

Surviving Covid: Resistance and Resilience of Movie Box Office Based on Machine Learning Prediction

YUCHENG FANG
FALL 2022

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Telecommunications and Media Industries and Mathematics
with honors in Telecommunications and Media Industries

Reviewed and approved* by the following:

Krishna Jayakar
Professor of Telecommunications and Media Industries
Thesis Supervisor

Rui Zhang
Assistant Professor of Computer Science and Engineering
Thesis Supervisor

Matthew Jackson
Associate Professor of Telecommunications and Media Industries
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

As one of the industries most impacted by the pandemic, the U.S./Canada movie box office market plummeted by 80% in 2020. However, anecdotal reports suggest that not all categories of movies were equally affected. This study focuses on investigating Covid's influence on different types of movies in the U.S. movie market, specifically how the movie box office of different movies changes under Covid. By utilizing a movies release database from SNL Kagan spanning 5 years, additional data from various sources, and machine learning methods, the study compares their actual box office receipts to the receipts predicted by a machine learning program. The predicted receipts are the movies' box office receipts if they were released in non-Covid times. The data shows that the overall movie budget is reduced in Covid and the return on investment (ROI) also declines significantly. In addition, the results show that PG-rated movies suffer the greatest loss in box office and R-rated movies have the least loss. The study proposes a way to inquire into different genres' resistance and resilience to Covid and theaters shutdowns.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Krishna Jayakar for his generous guidance, mentorship, and support throughout the process, without whom this thesis would be impossible. I would also like to thank Dr. Rui Zhang for his patient and enormous guidance on machine learning. I would also like to thank my honors advisers Dr. Jackson and advisor Del Schwab for their help during my study at Penn State. I would like to thank Dr. Bu Zhong for being my recommender to Schreyer Honors College.

I also would like to thank my mom not only for her constant love and support, but also for being a role model, a mentor, and a best friend to me.

Chapter 1

Introduction

The movie industry is an important part of the media and entertainment sector, including various means of distribution such as theatrical box-office, video rental and sales, pay television and increasingly, streaming media. The US box office market alone reached over 11 billion dollars in 2015. It is a major export revenue earner for the United States, as well as an important component of culture and soft power.

The movie industry is also highly risky due to the long production period and high production cost of movies. Yet the production cost is still growing. Of 943 movies that entered into production in 2021, 226 movies have an estimated budget greater than $15 million, which is a 124 percent increase from 2020 and a 40 percent increase compared to 2017 (THEME report, 2021).

Being commercially significant and risky, the movie industry attracts many researchers to predict movie success and study the industry using various methods since as early as 1980s (Litman, 1983; Litman and Ahn, 1998; Simonoff and Sparrow, 2000). These studies have used various methods including econometric methods and statistical models to explain which factors such as production investments, star power, MPAA rating, advertising and marketing, and release strategies affect the rate of return of movie investments. Interest in predicting movie success has continued more recently as well, with a number of recent papers investigating motion picture returns (Apala et al., 2013; Kim et al., 2014, 2020; Lash and Zhao, 2016;

Nihalaani et al., 2021; Quader et al., 2017; Ruus and Sharma, 2019; Zhang et al., 2008). However, research has not led to any consensus on factors predicting movie success.

Already confronting huge risks, the movie industry experienced a tremendous hit in terms of its box office market in 2020, when Covid happened and theaters had to shut down. The U.S./Canada theaters were fully closed for around 7 weeks in 2020, followed by partial openings in many parts of the market. Correspondingly, the U.S./Canada box office market dropped 80 percent from $11.4 billion in 2019 to $2.2 billion in 2020. The number is $4.5 billion in 2021, a 105 percent increase from 2020, but still remarkably lower than pre-Covid statistics (THEME report, 2021). The steep fall in box office revenue leads to the question of this study, how did different movie genres and rating categories behave under Covid? What types of movies suffered less loss or more? What movies would draw the audience into theaters despite the partial shutdowns and the health risks of going out in public during Covid?

To figure this out, this study examined how movies released during non-Covid times (i.e., those released pre-Covid or post-Covid), compared to those released during Covid. Non-Covid releases are used to train a predictive model of box office revenue, and the same model is applied to predict the revenues of Covid releases. In this way, the predicted revenue of Covid movies was estimated, assuming counterfactually that they were released during “normal” times. Then the study will compare the prediction to real box office by genre and MPAA rating to analyze resilience of different movies under Covid. In addition, the video units sold of Covid movies will be predicted to further investigate different movies’ behaviors.

Chapter 2

Literature Review

The movie industry is a high-profile and high-risk business because of the long production period and increasingly high marketing costs with high uncertainty of revenues. As one researcher found using 12 years of data on US theatrical releases, the association between costs and rates of return is not significant; although there is a positive association between production cost and revenues, the variance is high (Teti, 2013). Therefore, there are constant efforts in academia and industry to predict movie success. Although movies often generate other revenues, such as derivative cultural products and digital streaming revenues, the theatrical performance is the predictor and determinant of subsequent performance. Hence, most predictive studies focus on box office revenues.

In the 1990s, when movies became an increasingly important part of the entertainment sector, scholars attempted to uncover the factors underlying the financial success of motion movies using psychological approaches and economic approaches (Litman and Ahn, 1998). Econometric methods and statistical models were employed to explain different phenomena in the movie market (Litman and Ahn, 1998; Simonoff and Sparrow, 2000). Being the first to use multivariate regression model to predict movie box office, Litman repeated his "baseline" study with new data and updated variables to investigate the dynamics of the movie market (Litman, 1983; Litman and Kohl, 1989; Litman and Ahn, 1998). Simonoff and Sparrow (2000) used regression models to study the accuracy of models with different available variables. At this stage, direct audience measurement was usually absent and was replaced by other dependent variables (Litman and Ahn, 1998).

As more data has now become available online, machine learning techniques have emerged as an effective tool to predict the outcomes of movie releases. Athey and Imbens (2017) illustrated the effectiveness of machine learning in improving econometric methods in terms of causal effects. Liu and Xie (2019) also demonstrated that machine learning methods are superior in detecting data irregularities and therefore effective in short-term prediction. The accuracy of machine learning predictions usually depends on the number of items in the dataset and the extraction and organization of prediction variables. Accordingly, the starting point for this research is to identify a comprehensive list of predictors of motion picture success from the literature and organize them into groups. This paper divides the frequently considered predictors into 5 categories, which are movie-based, production-based, distribution-based, performance-based, and innovative predictors.

Movie-based predictors: Movie-based predictors are intrinsic features of the movie itself. This includes what the movie is, such as genre (Lash and Zhao, 2016; Quader et al., 2017; Zhang et al., 2008; Kim et al., 2014; Apala et al., 2013; Nihalaani et al., 2021; Simonoff and Sparrow, 2000; Kim et al., 2020), MPAA rating, sequel, runtime (Ruus and Sharma, 2019; Nihalaani et al., 2021), and plot synopsis (Lash and Zhao, 2016; Nihalaani et al., 2021). Most studies include genre and MPAA rating, which are basic and accessible, yet impactful on profitability. The plot synopsis is less accessible and requires data processing, and therefore is less frequently included.

Production-based predictors: The production-based variables are those involved in the production stage, such as movie budget, the director, the cast, and the studio. Budget is included in most studies since it has significant impact on the box office success of a movie. The influence of the cast and studio has been evaluated differently in studies. Most studies include star power but conceptualize it in different ways, including the first-billed actor or actress's tenure (Lash

and Zhao, 2016), gross or average revenues of the actor or actress's movies (Lash and Zhao, 2016; Quader et al., 2017), number of movies the actor or actress has been in (Simonoff and Sparrow, 2000), Google search records (Zhang et al., 2008), followers on Twitter (Apala et al., 2013), Facebook likes (Nihalaani et al., 2021), popularity polls and award nominations (Brewer et al.), and referring to relevant list (Kim et al., 2014). The director's star power is measured in similar ways to that of stars, by gross or average revenues of the director's movies (Lash and Zhao, 2016; Quader et al., 2017; Kim et al., 2014), followers on Twitter (Apala et al., 2013), Facebook likes (Nihalaani et al., 2021), and referring to relevant list (Kim et al., 2014). The studio power is captured by the number of film titles previously released by the studio (Kim et al., 2014).

Distribution-based predictors: Distribution-based variables highlight features involved in the distribution stage, such as marketing cost (Zhang et al., 2008; Kim et al., 2014), release season (Lash and Zhao, 2016; Quader et al., 2017; Zhang et al., 2008; Kim et al., 2014; Simonoff and Sparrow, 2000; Brewer et al., 2009), running days (Kim et al., 2014; Kim et al., 2020), number of screens (Quader et al., 2017; Zhang et al., 2008; Kim et al., 2014; Ruus and Sharma, 2019; Simonoff and Sparrow, 2000; Brewer et al., 2009; Kim et al., 2020), and competition (Zhang et al., 2008). Release seasons are defined differently: by weekend- and non-weekend releases (Kim at al., 2014), seasons (Lash and Zhao, 2016), months (Quader et al., 2017), and holiday releases (Lash and Zhao, 2016; Simonoff and Sparrow, 2000; Brewer at al., 2009). Although there is no universal definition of release seasons, a holiday release is considered to be different from others, because studios often reserve their best movies for holiday releases to capitalize on the bigger market during holidays. In effect, a holiday release could have a positive effect on movie box office, though competition is also higher during the holidays since all

studios tend to keep their best movies for this season. Some scholars have also aimed to measure competition independently of the season, by counting the number of movies released within 7 days of a movie's release date to evaluate initial competition (Lash and Zhao, 2016). This method effectively characterizes the effects of competition on movies for the first week of release, and is very important in predicting the eventual gross revenue since receipts during the opening weekend may create good news coverage and word-of-mouth publicity contributing to higher revenues in later weeks as well.

Performance-based predictors: Performance-based variables evaluate movies on the basis of audience reception, critical views, and awards. These features include the movie's audience ratings, critical evaluations, and awards. Many studies used the movie rating on fan sites and trade websites as the direct reflection of audience reception. Quader et al. (2017) included 6 different ratings from IMDb, Rotten Tomatoes, and Metacritic respectively. Similarly, ratings from these three websites are employed separately or combined in different studies (Nihalaani et al., 2021; Brewer et al., 2009). Some studies utilized sentiment analysis in their predictions to assess audience's response to the movie from textual analyses of IMDb and Rotten Tomato reviews (Quader et al., 2017), and YouTube comments on the movie trailers (Apala et al., 2013). Ruus and Sharma (2019) used Metacritic and the Indian fan site, SahiNahi.com for critical review for the movies. For audience reception, they calculated the number of tweets using hashtags for the movies. One of the early studies (Simonoff and Sparrow, 2000) used the ratings given by a famous film critic. For awards, Simonoff and Sparrow concluded that award nominations or wins benefited movies. But scholars tended to rely on a number of different awards, such as the Academy Awards (Simonoff and Sparrow, 2000; Brewer et al, 2009), the

Golden Globes, the (Orange) British Academy of Film Awards and the Screen Actors Guild Awards (Brewer et al, 2009).

Innovative predictors: Some studies created novel metrics through feature engineering and data extraction to evaluate invisible factors that contribute to a movie's success. When evaluating the effects of the cast on movie, Lash and Zhao (2016) not only used available data, but also engineered network-based features to assess diversity in the cast team, collaborations among actors, and collaborations between actors and directors. They also created hybrid features to describe the dynamics between different features, such as how the actors' previous experiences with a specific genre would influence the movies' revenue.

Overall, there are two types of data. One is the type that can be retrieved from reliable sources, such as genre, run time, and peak sites. The other is the type that configured differently in different studies, such as star power, director power, and competition. For this type of data, there is no commonly accepted definition or calculation. The evaluations usually depend on scholar's understanding and judgement of movie industry.

There are two definitions are often used by scholars to define movie success. They are box office revenue (Quader et al., 2017; Zhang et al., 2008; Brewer et al.,2009) and return on investment (Lash and Zhao, 2016; Kim et al., 2022), which is calculated as follows:

$$ROI = (revenue - budget) / budget$$

Because of the high standard deviation of movie box office revenues and production costs as shown in table 3, some believe ROI is a more meaningful way to evaluate movie success in terms of profitability because it considers production costs (Lash and Zhao, 2016; Kim et al., 2022). However, in this study, the production cost of movies in Covid is significantly lower than it in non-Covid period as shown in table 4. Therefore, ROI may not give as much information as movie box office revenue. The

movie success is measured by total theater revenue of the movie, which is the sum of Cumulative Box Office Gross Revenue and Theatrical Rental Revenue.

Chapter 3

Methodology

As stated in the introduction, this study examines how closely (or not) a predictive model of box office revenue, trained using data from "normal" times, is able to predict the box office returns of movies released during the Covid pandemic. The overall goal is to analyze the resilience of different movie genres and ratings categories under Covid. This chapter describes the data and methods used for the study.

3.1 Data Collection

The first phase is data collection. The main sources of data are SNL Technology, Media & Telecommunications (formerly SNL Kagan Unlimited) datasets, and two complementary data sources: IMDb and Box Office Mojo. Two comprehensive datasets are from SNL database, which are the Film Release Report and the Video Release Report. The Film Release Report includes the most important data for movies, which are film Genre, Studio, MPAA Rating, Theatrical Release Date, Film Peak Sites, Negative Cost, Print and Ad Cost, Cumulative Box Office Gross Revenue, and Theatrical Rental Revenue. The Video Release Report is data relating to the video market, including Video Units Sold in the US, Average Video Wholesale Price, Video Revenue, and Theatrical to Video Release Window. The Video Release Report covers fewer movies than The Film Release Report because not all movies are released in the video market. However, it is inaccurate to say that a movie not included in the Video Release Report was not released on video at all. Therefore, it is more accurately coded as "missing data," rather than as zero video release revenues.

A total of 3,243 movies released from May 2016 to August 2022 are retrieved from the SNL database and two reports are jointly used. Other data important to movie success prediction are collected from IMDb and Box Office Mojo. The Metascore (a weighted average of the scores awarded to movies by reputed critics) and awards are retrieved from IMDb. The sequel movies information is obtained from an IMDb list: Movies with sequels, which records all movies with sequels. In-release days and running time are retrieved from Box Office Mojo. For data from these websites, the author of this thesis developed web scrapers in the Python, and used it to parse HTML data from web pages and retrieve the data

3.2 Data Cleaning

In the second phase, data from all sources is first cleaned, processed, and consolidated into a single database. In this phase, data is processed so that it is ready to fit in the model. For example, data scraped from websites such as awards and in-release days are text-based information and numeric information needs to be extracted from it. Features like release season, studio, and foreign source are coded in this phase. In the end, the dataset is divided into two datasets, a non-Covid releases dataset (including pre-Covid and post-Covid releases) and a Covid releases dataset.

3.3 Feature Extraction

As most studies did, basic features such as Genre, MPAA rating, Total Cost and Cumulative Box Office Revenue are reported directly in the SNL datasets. The model also employs the variable “Foreign Film” to identify whether the movie is from the US or from

abroad because research has shown that US audiences tend to prefer US movies due to reasons like culture familiarity. In fact, research has shown that local audiences tend to prefer movies made in their country, to those made abroad; in other words, there is a "cultural discount" on foreign film investments (Jayakar and Waterman, 2000).

Studio Power is evaluated in two ways. One is the number of films the studio produces in the year the movie is released. The other is a binary variable to distinguish the movie produced by the Big Five, Universal Pictures, Paramount Pictures, Warner Bros., Walt Disney Pictures, and Columbia Pictures. These major studios are differentiated from other studios because of their strong influence on distribution networks, their high-quality movies, and the high investment. While star power and director power are popular predictors in the literature and are believed to have influence on movie success, they are not included in this model. The reason is that the dataset used in this study spans from 2016 to 2022, which makes it difficult to access star power or director power at different times.

For release season, there is no commonly recognized definition, with various authors differentiating between weekend and non-weekend releases, as well as by month and holiday release. As this study focuses on the US market, it uses Christmas release and summer release as two important movie release seasons. The competition is measured by the number of films released in 7 days after the movie is released because a movie's performance in the opening week is critical to its success.

Metascore and awards are employed to reflect critical and professional evaluations of the movie. The Awards is counted by adding all nominations and wins from global film festivals and awards. Therefore, the range of the number of awards goes from 0 up to over 500, such as Parasite, the Korean movie which achieved 579 wins and nominations globally. Total 15 features

are used to feed the model to predict movie revenues and 16 features are used to predict video unites sold in the US. Table1 shows the summary of all features, its definitions, and source.

A number of variables related to the video-release are available in the SNL database. However, they are not used in the prediction because the box office release pre-dates video release in the industry's traditional release window sequencing. IMDb rating is not used for the same reason.

Table 1 Variables, their definitions, and sources

Variable	Definition	Data Source
Genre	Genre of the film	SNL Database
MPAA Rating	MPAA's ratings for films	SNL Database
Sequel	The movie continues a story begun in a previous movie	SNL Database
Runtime	The length of the film	Box Office Mojo
Studio	Film studio name	SNL Database
Negative cost	Direct costs related to the creation of a film	SNL Database
Print and ad cost	Costs related to the creation of film prints and advertising for a film within the U.S.	SNL Database
Foreign	The film is foreign if it is not a U.S. film.	Feature Engineering
Total cost	Negative cost + print and ad cost	Feature Engineering
Studio Power	The number of films the studio produces in the year the film is released	Feature Engineering
Release season	The film released in December is identified as Christmas release. The film released in July or August is identified as summer release.	Feature Engineering
In-Release days	The number of days that the film is on screen	Box Office Mojo
Peak sites	Peak number of sites at which a film has been shown	SNL Database
Competition	The number of films released in 7days of the film's release date.	Feature Engineering
MetaScore	A score created by assigning scores to their reviews of a large group of the world's most respected critics, and a weighted average is applied to summarize the range of their opinions.	IMDb
Awards	The sum of all nominations and wins	IMDb
Cumulative Box Office Gross Revenue: U.S. ($000)	Cumulative box office gross revenue generated by a film in the U.S. release	SNL Database
Theatrical Rental Revenue: U.S. ($000)	Revenue generated from the fees theaters pay to rent a film's print for U.S. exhibition	SNL Database
Video Units Sold: U.S.	Total units of a film sold in any video format in the U.S.	SNL Database
Video Revenue: U.S. ($000)	Total revenue from all physical video media sales in the U.S.	SNL Database

3.4 Summary of Descriptive Statistics

There are 3,243 movies released in the US between May 2016 to August 2022 in the original dataset. Some movies are repeated because they have two theatrical release dates. After removing these duplicated movies, there are 3,236 movies in the final dataset. Not all movies have complete information. Information may be unavailable for a variety of reasons. For example, some movies do not have video-related information because the video of the movie is not released. Another reason can be that the information is missing and thus unretrievable from IMDb or Box Office Mojo websites, such as running time. In the case of running time, missing values are replaced by the mean value of the variable to enable observations with missing values to be used in the rest of the analysis. However, for variables that are important predictors of movie success, such as Metascore, replacement with the average is not a valid method. Therefore, these variables are coded as –1 for the model to eliminate from the analysis. Table 2 shows the variables with the number of observations for which data are available, and the percentage availability.

Table 3 shows the summary financial statistics of movies in the final dataset. Movies produced in the 6 years for which data are collected in this study cost $16 million on average to produce, though there was a wide variation, with the cheapest costing just $5,000 and the most expensive $361 million. Pint and advertising costs were 40-50% of negative costs, testifying to the importance of marketing and promotion in the movie business. Average cumulative box office was just $1.5 million, indicating that a large number of movies failed to recoup their production and marketing expenditures. However, the most successful movies gained enormous box office revenues (a maximum of $858 million), effectively compensating for the large number of smaller, loss-making films. Confirming the findings of the literature, it is clearly seen

that the movie industry is very risky as the Cumulative Box Office Gross Revenue has a rather high standard deviation, compared to the mean. Meanwhile, Theatrical Rental Revenues (representing the studios' share of box office revenue) had a higher mean, since the studios are compensated at higher rates for blockbuster films.

Table 2 Number of observations for which data are available and percentage availability

Variables	N	Available Percentage
Film Title	3236	100%
Film Genre	3236	100%
Studio	3236	100%
MPAA Rating	3236	100%
Theatrical Release Date	3236	100%
Film Peak Sites	3236	100%
Negative Cost ($000)	3236	100%
Print and Ad Cost: U.S. ($000)	3236	100%
Cumulative Box Office Gross Revenue: U.S. ($000)	3236	100%
Theatrical Rental Revenue: U.S. ($000)	3156	97%
In-release days	1141	35%
Earliest Physical Video Release Date	1403	43%
Video Units Sold: U.S.	1403	43%
Average Video Wholesale Price ($)	1403	43%
Video Revenue: U.S. ($000)	1403	43%
Video Revenue: U.S./ Box Office: U.S. (%)	1403	43%
Theatrical to Video Release Window	1403	43%
runningtime_min	1535	47%
metascore	2266	70%
awards	3236	100%

Though video release revenues also displayed a high standard deviation relative to the mean, the magnitude of the difference is much smaller relative to the theatrical box office. In other words, video releases are less risky than theatrical releases. This is because the box office release experience acts as a filter, and non-performing movies are not released to video at all.

The survivors that do appear in video release tend to be on average less risky, and therefore with lower standard deviation in revenue.

Table 3 Summary statistics for key financial variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Cumulative Box Office Gross Revenue: U.S. ($000)	1599	57605	0.024	858366
Negative Cost ($000)	16008	35961	5	361233
Print and Ad Cost: U.S. ($000)	9013	17168	11	96622
Theatrical Rental Revenue: U.S. ($000)	8393	30518	0	46987
Video Units Sold: U.S.	233157		1000	5630000
Video Revenue: U.S. ($000)	4628	13057	0	120116
Metascore	63	16	1	100

To further investigate the data, the dataset is divided into two datasets by time: non-Covid release dataset and the Covid release dataset. The Covid timelines and polices are different country by country and there is no precise definition of its start and its end. Since this study focuses on the US movie market, it uses the Center for Disease Control's (CDC) Covid-19 timeline as reference. March 15, 2020, is marked as the beginning of Covid, when states began to implement shutdowns in order to prevent the spread of COVID-19 and the New York City public school system, the largest school system in the U.S., shut down. The end is marked by April 1, 2022, when the CDC announced the termination of Title 42, an Order that suspended the admission of migrants into the U.S. due to the public health risk of COVID-19. In addition, April 2022 is when movie theaters were fully reopened (CBS News). Therefore, the movies released between March 15, 2020, and April 1, 2022, are divided into the "Covid release" dataset, and the rest of the movies are in the non-Covid release dataset.

The non-Covid release dataset contains 2,985 entries and the Covid dataset contains 250 entries. The summary statistics of two datasets are shown in Table 4. For revenues, both

Cumulative Box Office Gross Revenue and Theatrical rental revenue have significantly lower means during Covid as expected due to the shutdowns of theaters. The Negative Cost, which is the production cost, and Print and Ad Cost also decrease remarkably during Covid. The change is reasonable since production companies may want to decrease the production cost or put off big productions as the box office is predictably worse than in non-Covid times. The Video Units Sold in Covid times is lower than in non-Covid times; this is surprising since the closure of theaters might be expected to create additional demand for videos and DVDs, close substitutes for theatrical exhibition. This might be explained by the closure of many physical stores during Covid, as well as by the growing popularity of streaming platforms and the long-term decline in video rentals and sales.

Table 4 Comparing non-Covid releases and Covid releases

	Mean		Standard Deviation		Minimum		Maximum	
Variable	Covid	Non-Covid	Covid	Non-Covid	Covid	Non-Covid	Covid	Non-Covid
Cumulative Box Office Gross Revenue: U.S. ($000)	2273	17047	10275	59773	0.256	0.024	100392	858366
Negative Cost ($000)	10696	16458	25994	36647	25	5	207788	361233
Print and Ad Cost: U.S. ($000)	3421	9484	7158	17675	57	11	50556	96622
Theatrical Rental Revenue: U.S. ($000)	1458	8992	5382	31700	0	0	51200	469870
Theatrical to Video Release Window	111	125	76	107	4	3	550	1104
Metascore	60	63	16	16	14	1	95	100

For the video units prediction model, the missing values of Video Units Sold are taken away from both non-Covid and Covid datasets, which left 1266 entries in the non-Covid dataset and 137 entries in the Covid dataset. Then the movie box office revenue is added to predict video units sold.

3.5 Machine Learning Algorithms and Evaluation

The main purpose of the study is not to test models on predicting the US movie market. Many previous studies concentrate on optimizing models to their best performances rather than interpreting the findings. The purpose is to use machine learning as an effective tool to analyze the impact of Covid on the movie market by comparing its predictions to actual outcomes. Therefore, the machine learning algorithm used in the study is decision trees. Decision trees algorithm has better explainability and interpretability compared to other models. It can be visualized and used to explicitly represent the decision-making process. Since box office revenue is used to evaluate movie success, the MSE (mean-square error) and R2 will be used as performance metrics to evaluate the models.

Two models are fed to assess Covid's impact on movie markets, one is on movie revenue (model 1), and the other is on video units sold (model 2). First, the non-Covid dataset is used to train the model. The non-Covid data is split into two: 70% of the data is used as training data and the rest 30% is used as testing data. The maximum depth of the tree is limited to 10. The minimum number of samples required to split is 30 and the minimum number of samples required to be at a leaf node is 10. These parameters are set to avoid overfitting. Then, with the model trained, the Covid dataset was fed into the model to predict box office revenues of movies released in Covid. The results are expected box office revenues of these movies if they were counterfactually released in non-Covid times. The results are analyzed to investigate and explain Covid's influence on different movie genres and rating categories in terms of box office revenues.

Chapter 4

Results

4.1 Budget and Movie Revenue

In Figure 1 and Figure 2, the relationships between production budget and box office revenue in Covid times and non-Covid times are shown respectively. The scatterplots show that there is a strong positive and linear association between production budget and box office revenue. However, both production budget and box office revenue in Covid times are significantly lower than in Covid times. The decrease in the production budget is reasonable as investors have lower expectations of their payback during Covid and thus decrease their investment in movies. Figure3 shows the regression lines of Covid movies and non-Covid movies. It demonstrated that the revenue-cost ratio of Covid movies is remarkably lower than non-Covid movies.

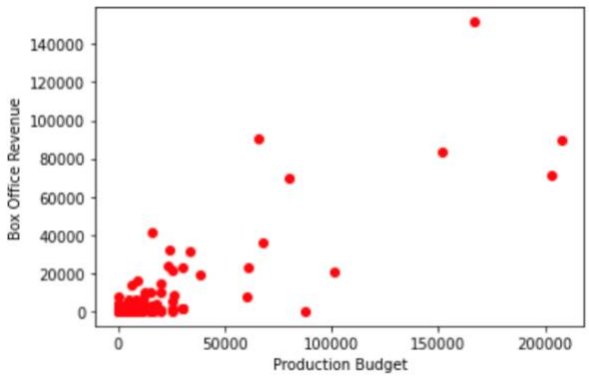

Figure 1 Scatter plots of production budgets and box office revenues for movies in Covid period

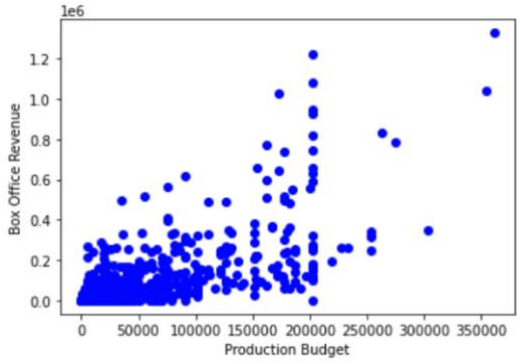

Figure 3 Scatter plots of production budgets and box office revenues for movies in non-Covid period

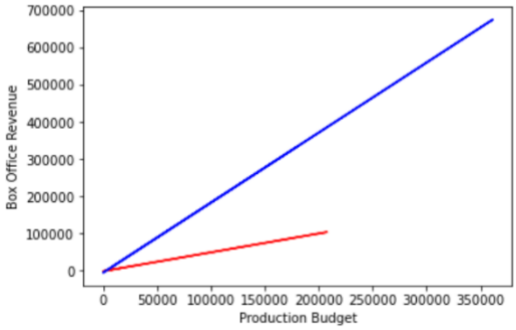

Figure 2 Trend lines for Covid (red) and non-Covid (blue) periods

4.2 Gini Coefficient

The study also employs Gini coefficient to measure the inequality of the movie market. The Gini coefficient is calculated by the area between the line of perfect equality and the Lorenz curve, which plots the Y proportion of total income that is cumulatively earned by bottom X% of the movies, over the area under the line of perfect equality. Figure 4 and Figure 5 show the Gini coefficient and Lorenz curve of movie revenue of Covid times and non-Covid times respectively. Surprisingly, the Gini coefficient of Covid movies is only slightly higher than that of non-Covid movies, which means the inequality of movie revenue is barely strengthened during Covid.

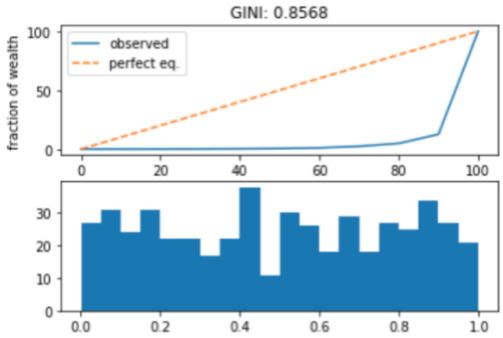

Figure 4 Gini coefficient and Lorenz curve for movies in Covid period

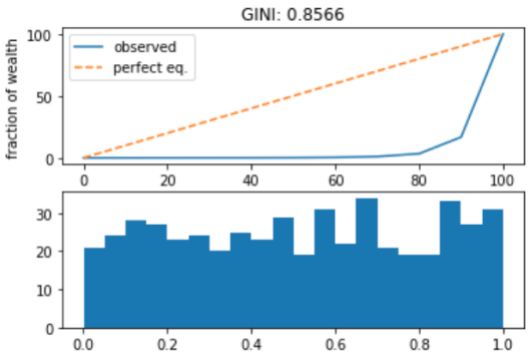

Figure 5 Gini coefficient and Lorenz curve for movies in non-Covid period

In combination, the scatter plots and Gini curves appear to indicate that though movies lost revenue during Covid, the contraction equally impacted all movies, with the result that the distribution of revenues was almost identical to non-Covid times. The next section examines this preliminary finding in more detail by examining the difference between the revenues predicted by the machine learning program and the actual revenues for movies during Covid.

4.3 Model Results and Analysis

Model 1 gives the prediction of revenues, which is the expected revenues of Covid movies if they are released in non-Covid times. The change of revenues is measured by:

(actual revenue - predicted revenue) / real revenue

Table 5 compares revenue change by genre and MPAA rating. Thus, the numbers in Table 5 indicate how much higher the predicted revenue would have been, had the movie been released in a non-Covid environment. For example, a PG-13 Romance movie would have received 1.88% more revenue during non-Covid times, compared to its actual revenue. In other words, the numbers represent the lost percentage revenue due to Covid for each genre-rating combination. NA represents no entries falls under the category. For example, there is no G rated Romance movie released during Covid. NR rating is removed as Not Rated movies give limited information. As shown in the table, Romance, Horror, and Thriller movies experience less but still substantial loss during Covid. Foreign and Documentary movies suffer heavy loss in revenue. In terms of MPAA rating, R-rated movies have the least loss and G-rated and PG-rated movies have more loss.

Table 5 Percentage change in real revenue compared to predicted revenue

Film Genre	MPAA Rating				
	G	PG	PG-13	R	All
Romance	NA	NA	-1.88	NA	-1.88
Horror	NA	NA	-0.81	-4.02	-3.66
Thriller	NA	NA	-0.37	-0.19	-3.69
Comedy	NA	-4.71	-1.88	-3.57	-5.09
Sci-Fi/Fantasy	NA	-6.17	NA	NA	-6.17
Western	NA	NA	-8.19	NA	-8.19
Drama	NA	-33.11	-6.29	-3.81	-9.08
Family	-27.18	-7.07	NA	NA	-13.77
Action	NA	NA	-9.61	-9.90	-13.83
Animated	-18.78	-20.17	NA	NA	-18.91
Foreign	NA	-6.14	-4.64	-5.96	-30.60
Documentary	NA	-223.246	-20.53	NA	-63.87
All	-20.46	-34.53	-7.15	-4.60	-17.36

Model 2 gives the prediction of video units sold. The percentage change is calculated in the same way as of revenue:

(actual units - predicted units) / real units

There is no G-rated movie video released in Covid, therefore the column is removed. In terms of genre, video units sold of Animated, Action, Thriller, and Family movies achieve growth during Covid. Comedy encounters a great decrease in video units sold, which is also the only below average change among all genres. For MPAA rating, the percentage change of video units sold sees an upward trend from PG to R.

Table 6 Percentage change in video units sold compared to predicted units sold

Film Genre	MPAA Rating			
	PG	PG-13	R	All
Animated	0.76	NA	NA	0.76
Action	NA	0.44	0.34	0.30
Thriller	NA	0.36	0.32	0.10
Family	0.03	NA	NA	0.03
Horror	NA	-0.62	-0.033	-0.01
Documentary	0.84	NA	NA	-0.10
Foreign	0.53	0.02	-0.49	-0.03
Romance	NA	-0.60	NA	-0.60
Western	NA	-0.64	NA	-0.64
Drama	NA	-0.92	-1.08	-0.84
Comedy	-0.34	-25.42	-2.05	-7.88
All	0.59	-3.41	-0.67	-1.06

Chapter 5

Conclusions and Future Work

The study proposes a method to study Covid's influence on the movie market and test different movies' resilience to Covid by genre and MPAA rating. To achieve the goal, the study tries to predict movie success in terms of box office revenues based on various features from different sources. The data of non-Covid movies are used to train the model with decision tree algorithm to predict revenues of non-Covid movies in non-Covid times. The same is done for video units sold to check if a reverse pattern happens to video market since if movies of a certain category suffer huge loss in box office revenue that watching them in theatres is not as attractive as watching other movies, they are expected to sell more video units.

The study proposes a method to study Covid's influence on the movie market and test different movies' resilience to Covid by genre and MPAA rating. To achieve the goal, the study tries to predict movie success in terms of box office revenues based on various features from different sources. The data of non-Covid movies are used to train the model with decision tree algorithm to predict revenues of non-Covid movies in non-Covid times. The same is done for video units sold to check if a reverse pattern happens to video market since theatrical exhibition and video rentals and sales are economic substitutes. If movies of a certain category suffer huge loss in box office revenue that watching them in theatres is not as attractive as watching other movies, they are expected to sell more video units.

This study finds that in terms of revenue, G-rated and PG-rated movies are greatly influenced by Covid, and PG-13-rated and R-rated movies are less influenced, although PG-13

and R have smaller audience base. This result can be explained by the assumption that theatres and big screens are able to showcase the special effects and production values of PG-13-rated and R-rated movies, and therefore the audience are more likely to go to theaters for those movies than G-rated and PG-rated movies in Covid. The video market result displays a different pattern that PG-rated movies have the best performance and PG-13-rated and R-rated movies have inferior performance.

This study can be helpful to understand and interpret the effect of audience choices on different movie genres and rating categories during Covid. It is informative for investors to evaluate their investment during Covid. Although the movie market as well as the world is recovering from Covid, this study is helpful for any stakeholder who cares about the success of movies to make strategies in the case of similar events, for example, what kind of movies to release in theaters or what kind of movies to invest in. Furthermore, the study is under the background of theaters shutdowns, hence provides reference for streaming platforms and movie practitioners about what movies are more attractive on big screens or home devices.

There is a lot to be improved about this study. One of the directions is to optimize the predictive models. For example, streaming revenue is an important part of revenue during Covid. However, it is not included in this study because the data is unavailable. More analysis can be done with information on how different movies perform in streaming platforms. In addition, the prediction can be improved by using more algorithms and choosing the best after comparing the results.

BIBLIOGRAPHY

Apala, K. R., Jose, M., Motnam, S., Chan, C.-C., Liszka, K. J., & de Gregorio, F. (2013). Prediction of movies box office performance using social media. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. https://doi.org/10.1145/2492517.2500232

Athey, S., & Imbens, G. W. (2017). The state of Applied Econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Brewer, S. M., Kelley, J. M., & Jozefowicz, J. J. (2009). A blueprint for success in the US Film Industry. *Applied Economics*, *41*(5), 589–606. https://doi.org/10.1080/00036840601007351

Jayakar, K., & Waterman, D. (2000). The economics of American theatrical movie exports: An empirical analysis. *Journal of Media Economics*, 13 (3), 153-169.

Kim, J.-M., Xia, L., Kim, I., Lee, S., & Lee, K.-H. (2020). Finding nemo: Predicting movie performances by Machine Learning Methods. *Journal of Risk and Financial Management*, *13*(5), 93. https://doi.org/10.3390/jrfm13050093

Kim, T., Hong, J., & Kang, P. (2014). Box Office forecasting using machine learning algorithms based on SNS Data. *International Journal of Forecasting*, *31*(2), 364–390. https://doi.org/10.1016/j.ijforecast.2014.05.006

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

Litman, B. R., & Ahn, H. (1998). Predicting financial success ofmotion pictures: The early ’90s experience. In B. R. Litman (Ed.), *Motion picture mega-industry* (pp. 172–197). Needham Heights, MA: Allyn & Bacon.

Litman, B. R., & Kohl, L. S. (1989). Predicting financial success of Motion Pictures: The '80s experience. *Journal of Media Economics*, *2*(2), 35–50. https://doi.org/10.1080/08997768909358184

Liu, Y., & Xie, T. (2019). Machine learning versus econometrics: Prediction of box office. *Applied Economics Letters*, *26*(2), 124–130. https://doi.org/10.1080/13504851.2018.1441499

Motion Picture Accociation. (2022). *THEME Report 2021*. Retrieved November 8, 2022, from https://www.motionpictures.org/wp-content/uploads/2022/03/MPA-2021-THEME-Report-FINAL.pdf

Nihalaani, R., Shete, A., & Khan, D. (2021). Movie success prediction using naïve Bayes, logistic regression and Support Vector Machine. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. https://doi.org/10.1109/icrito51393.2021.9596138

Quader, N., Gani, M. O., Chaki, D., & Ali, M. H. (2017). A machine learning approach to predict movie box-office success. *2017 20th International Conference of Computer and Information Technology (ICCIT)*. https://doi.org/10.1109/iccitechn.2017.8281839

Ruus, R., & Sharma, R. (2019). Predicting movies' box office result - A large scale study across Hollywood and Bollywood. *Complex Networks and Their Applications VIII*, 982–994. https://doi.org/10.1007/978-3-030-36683-4_78

Simonoff, J. S., & Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *CHANCE*, *13*(3), 15–24. https://doi.org/10.1080/09332480.2000.10542216

Teti, E. (2013). The Dark Side of the movie. the difficult balance between risk and return. *Management Decision*, *51*(4), 730–741. https://doi.org/10.1108/00251741311326536

Zhang, L., Luo, J., & Yang, S. (2008). Forecasting box office revenue of Movies with BP neural network. *Expert Systems with Applications*, *36*(3), 6580–6587. https://doi.org/10.1016/j.eswa.2008.0

YUCHENG FANG

EDUCATION

The Pennsylvania State University, Schreyer Honors College — **University Park, US**
- *BA in Mathematics & Telecommunications and Media Industries* — *Aug 2019 – Dec 2022*
- **Honors**: Evan Pugh Scholar Award (2022); The President's Freshmen Award (2019)
  Dean's List for all semesters

Jinan University — **Guangzhou, CN**
- *Majoring in Advertising* — *Sep 2017 – Jun 2019*

London School of Economics and Political Science — **London, UK**
- *Summer Program in Marketing* — *Jul 2018 – Aug 2018*

INTERNSHIP EXPERIENCE

Product Strategy & Operation Intern, **ByteDance Ltd, Beijing, CN** — *Jan 2021 – May 2021*
- Performed routine debugging and provided insights for product optimization, such as external chain verification and Emoji user-side differentiation mechanism
- Followed up the go-to-market process of new data metrics, designed and conducted beta tests in collaboration with platform PSOs, and produced the data logic checklist together with the PM for local PSOs
- Retrieved data from internal BI, analyzed relevant metrics using Excel, proved the combined purchase of hashtag challenge (HTC) and brand effect (BE) incurring a higher return rate than a single product, and visualized the findings into one-pager for local PSOs' promotion of products; Classified all the HTC and produced relevance index definition roadmap

Marketing Intern, **New Leaf Initiative**, **State College, US** — *Jan 2020 – Mar 2020*
- Operated Facebook account, engaged followers in online interactions, and sustained follower community via social media
- Designed promotional posters using Canva and Photoshop for social media posts and offline activities and assisted in networking events to strengthen community connection

Marketing & Public Relations Intern, **Tencent Holdings Ltd, Guangzhou, CN** — *Feb 2019 – Jul 2019*
- Established social media matrix, produced two circulation posts with over 200,000 reads, attracted over 250 followers on TouTiao in one month; became the hottest account in the number of followers and reading volume
- Assisted in on-site campaigns at Jinan University and attracted over 200 students to participate in the activities
- Planed promoting activities, selected topics, drafted tweets (average reads around 1000), and developed operation strategies on Zhihu as a core member of the "Xplore Prize" WeChat public account

Account Executive Intern, **Shanghai Ogilvy & Mather Advertising Co., Ltd**, **Guangzhou, CN** — *Mar 2019 – Jul 2019*
- Launched campaigns on media platforms including Google, Facebook, YouTube, and Snapchat, screened the most cost-effective keywords, and maximized the click rate through real-time monitoring
- Researched the social media marketing of clients' competitive products in overseas markets in terms of TVCs, print, and campaigns to examine their marketing strategies and provide clients with targeted marketing plans
- Sorted out media resources of over 10 countries with Excel; investigated media circulation, influence and advertising price

ACADEMIC PROJECTS

Surviving Covid Resistance and Resilience of Movie Box Office Based on Machine Learning Prediction
Honors Thesis, Advisor: Dr. Krishna Jayakar & Prof. Rui Zhang — *Aug 2021 – Dec 2022*
- Proposed an innovative method to study Covid's impacts on different types of movies by comparing their actual box office and predicted box office
- Built web scrapper using python to collect descriptive movie information from IMDb and Mojo Box Office
- Performed data cleaning and feature engineering with python to prepare datasets for the predictive model
- Developing predictive models using decision trees to predict the film box office data during COVID-19

Population-based HIV Impact Assessment (PHIA), *PSU URA Summer Project* — *Jun 2022 – Jul 2022*
- Analyzed the PHIA Data for a research project which is focused on the prediction of key metrics for subpopulations, in particular, Female Sex Workers in Dr. Le Bao's statistics research lab
- Assisted in developing methods to address missing data resulting from survey skip conditions

The Comparison between Public Broadcasting System in China&Singapore, *COMM419 Honors* — *Aug 2021– Dec 2021*
- Analyzed public broadcasting systems in the Asian context and compared public broadcasting systems in China and Singapore in terms of regulations and ownership
- Completed a 15-page academic paper of over 3800 words and presented the findings in class

Analysis of Tencent's Company Strategy, *Online project, Advisor: Prof. Paul Hardart (NYU)* — *Jul 2020 – Aug 2020*
- Learned about the evolution and development trend of mass media, explored the innovation of social media brought by new technology, and developed unique insights on Internet influencers and media effects
- Co-authored "A Comprehensive Analysis of Tencent's Company Strategy" and published in the proceedings of the 2020 11th International Conference on Economics, Business and Management (ICEBM 2020)

ON-CAMPUS ACTIVITIES

Event Planner, Math Club, PSU, Aug 2022 – Dec 2022
Student Ambassador, Student Ambassadors to Global Alumni (SAGA), PSU, Jan 2022 – Dec 2022
Actress, PSU Chinese Drama Society, PSU, Sep 2021 – Dec 2022
Teaching Assistant for COMM419 International Media System & COMM403 Mass Media Law, PSU, 2022 & 2020

ADDITIONAL

Computer: Python; SPSS; C++; R; Premiere; SQL
Certificates: Google Analytics Individual Certification (Oct 2022)
Coursera: Supervised Machine Learning: Regression and Classification, Stanford University (Sep 2022)