

Saving Cinema: Designing Movie Theater Showing Schedules

Project Proposal

Subject Area: Film Industry

Jeewon Han, Lucas Rozendaal, Jay Lawrence, and Jillayne Clarke

EN.553.602: Research and Design in Applied Mathematics: Data Mining

September 7, 2023

### **Proposed Research**

In recent years, the news media has frequently reported on the financial troubles of the movie theater industry in the aftermath of the COVID-19 pandemic. In 2020, A large number of movie theaters closed, and most major film releases were postponed during the height of the pandemic. According to data from The Numbers, the industry suffered an immediate, stark downturn and has been slow to recover in the years that followed. The 2019 U.S. box office was approximately \$1.99 billion, down from over \$11.23 billion the previous year, a decline of over 82 percent. As theaters reopened and the U.S. economy began to recover from the pandemic, domestic ticket sales and box office numbers have increased year-by-year. However, the annualized domestic box office for 2023 is only projected to be \$9.64 billion, still well below the domestic box office in 2019. Ticket sales have been similarly slow to recover: the annualized estimate of ticket sales in 2023 is still slightly less than 75% of what it was in 2019. Furthermore, the ongoing Writers Guild and SAG-AFTRA strikes are likely to put additional pressure on the movie theater industry as many highly anticipated upcoming movies push back their release dates or halt production (Lenker). To handle these challenging financial pressures, cinemas have to now be more pragmatic than ever when deciding how to structure their business and maximize their profit.

As such, our business question is, “How can a movie theater company decide the showing schedule for its theaters? In other words, which upcoming movie releases should the company show, how frequently should a given movie be run at the theaters, and for how long should that movie remain in the schedule?” To answer this question, we want to build a model using movie metadata to predict how profitable showing an upcoming film is likely to be for a movie theater company.

To begin building our model, we searched for a dataset containing various metadata on a large number of films. That is, we were looking for a movie dataset with features such as genre, director, cast, language, and, most importantly, financial information like budget and box office revenue. The most popular source of such information is IMDb (Internet Movie Database). Unfortunately, access to financial data from IMDb is restricted for commercial purposes. The free IMDb datasets for personal and academic use lack that information and would therefore limit the depth of our analysis if used. Instead, we plan to extract our data from TMDB (The Movie Database), a similar, lesser-known, user-editable database. TMDB's API is free and well-documented. More importantly, it has the financial data that we are unable to access through IMDb.

Using the TMDB data, we plan to craft a model that predicts a given movie's profitability for a movie theater company. We anticipate this will be done through some form of linear or nonlinear regression using what we determine to be key features from the dataset. We'll then compare our model against a baseline model, which will either be a simple model we build ourselves or potentially just consider a mean value for the target variable.

To better predict box office revenue, we hope to supplement the TMDB data with other information that may predict box office revenue, like data on opening weekend revenue. Our goal is not to predict a movie's box office revenue in general, however, but the profitability of showing that movie for a specific movie theater company. As such, we will need to look for data on revenue streams for movie theaters and how they relate to showings to create our model. We plan to reach out to the Warehouse Cinemas, a local theater at the Rotunda serving the Charles Village community, to see what insights we can glean.

### Data Explanation

TMDB (The Movie Database) is a user-editable database for movies and TV shows. It offers a free API that provides access to the data for hundreds of thousands of movies. As of this proposal, we have not yet had the time to call or clean the dataset from the API that we will be using in our final model. However, a much smaller subset of TMDB data from 2017 is available on Kaggle for our reference. Analyzing that smaller dataset enables us to make initial observations about the kind of information we will have access to. The size of the sample dataset is 4,803 movies and includes the features we are looking to incorporate in our model (as outlined in Section II). For a more thorough breakdown of the features in the sample dataset, see the table below:

**Table 1.** Features of Sample TMDB Dataset

Feature Name	Description
budget	Estimate of the film's production budget
genres	List of the film's genres
id	TMDB ID
keywords	List of keywords for TMDB search
original_language	The film's original language
original_title	The film's original title
overview	The film's official description
popularity	TMDB's popularity metric, based on interaction with the film on the site
production_company	List of the film's production companies
Production_countries	List of the film's production countries
release_date	The film's release date
revenue	The film's reported worldwide box office revenue
runtime	The film's runtime (in minutes)
spoken_language	List of languages spoken in the film
status	Variable indicating if the film has been released
title	The film's title
vote_average	The film's average rating on TMDB based on user votes
vote_count	The amount of ratings for the film

For this project, we're assuming that people decide to go to see a given movie based on information about that movie available before it is released. In other words, our assumption is that features such as cast, budget, director, genre, etc. are correlated with viewership and therefore revenue for movie theater companies. Before we can run our regression model, we will have to perform careful feature engineering to maximize the predictive power of our data. For example, since the cast of a movie is a categorical variable with an enormous number of levels, we will need to create some measure to represent the "star power" of a given cast.

Before we can begin our feature engineering, we will have to perform careful data cleaning and preparation. Because TMDB is a smaller, user-editable database, it is probably more likely that there will be errors or improper inputs in the data we extract. Even within the sample dataset, it is easy to find entries that have bad inputs. "A Farewell to Arms," for example, is listed as having a budget of "\$4" and a total box office revenue of "\$25," numbers that are off by a factor of one million (Kaggle). For these data cleaning and preparation steps, we may need to cross-reference figures like this or try multiple queries for the same data, taking care to avoid spurious auto-placement of delimiters.

Another possibility for assessing the quality of the data is to cross-reference with movie metadata extracted elsewhere. For everything other than financial data, one option would be to use the publicly available IMDb dataset to validate the information from TMDB. Furthermore, we could potentially also extract IMDb ratings or even data from Rotten Tomatoes and add that to our dataset as additional features for our model.

### **Potential Impact**

The movie theater industry has yet to recover from the financial challenges imposed by the COVID-19 pandemic. By next year, the SAG-AFTRA and Writers Guild strikes will likely

place even more strain on movie theaters as the production of new films slows down. During a period where movie theaters are facing this dire financial situation, it is particularly important for them to design their showing schedules to maximize their revenue.

As mentioned in Section II, we intend to reach out to Warehouse Cinemas at the Rotunda for information on the state of the industry and insight into their business model. Warehouse Cinemas opened earlier this year, taking over the lease from CinéBistro at the Rotunda, an older movie theater that was forced to close in the aftermath of the pandemic (Baltimore Business Journal). CinéBistro's operator, CMX Cinemas, declared Chapter 11 bankruptcy following decreased revenue during the COVID-19 pandemic. As such, the outcome at CinéBistro represents a particularly vivid example of the industry's struggles.

If we can successfully develop a better way for theaters of deciding how to design their schedules, we may be able to provide a business solution that could help movie theater companies increase their revenue. If successful, our model could prove helpful for companies like Warehouse Cinemas during this difficult period for their business.

## References

- Baltimore Business Journal. (2020, December 10). *CineBistro in Hampden closes as operator emerges from bankruptcy*. WBAL-TV. <https://www.wbalv.com/article/cinebistro-baltimore-hampden-closes-operator-emerges-from-chapter-11-bankruptcy/34934620>
- Kaggle. (2019, February 6). *TMDB Box Office Prediction: Can you predict a movie's worldwide box office revenue? [Data set]*. Kaggle. <https://www.kaggle.com/competitions/tmdb-box-office-prediction/overview>
- Lenker, Maureen Lee. (2023, June 30). *SAG strike explained: How an actors' strike will impact TV and movies*. Entertainment Weekly. <https://ew.com/movies/sag-actors-strike-everything-to-know-movies-tv/>
- The Numbers. *Domestic Movie Theatrical Market Summary 1995 to 2023*. The Numbers. <https://www.the-numbers.com/market/>

## Appendix – ChatGPT Inquiries

I have to create a data mining project based on the following guidelines: We are to be answering a business question by using data to analyze potential solutions. Can you help me generate some ideas with possible data sources (datasets) for projects we could pursue? The data should most likely be publicly available. Give us some analysis or modeling ideas per project as well. Some given examples include filtering fake news, detecting credit card fraud, and coming up with optimal airport hubs for a spoke-and-hub model for Southwest.

### ChatGPT

Certainly! Here are some data mining project ideas along with possible data sources and analysis/modeling ideas for each:

1. **Customer Churn Prediction for a Telecom Company:**
  - **Data Source:** Publicly available telecom customer data, including historical customer information, call records, billing data, customer service interactions.
  - **Analysis/Modeling Ideas:** Build predictive models to identify customers at risk of churning. Explore factors like customer demographics, call patterns, contract details, and customer service interactions. Use machine learning algorithms like logistic regression, decision trees, or random forests.
2. **Retail Store Sales Forecasting:**
  - **Data Source:** Publicly available sales and inventory data for a retail store chain.
  - **Analysis/Modeling Ideas:** Develop time series forecasting models to predict future sales. Explore factors like seasonality, holidays, promotions, and economic indicators. Methods such as ARIMA, exponential smoothing, or machine learning-based models like LSTM can be used.
3. **Stock Market Price Prediction:**
  - **Data Source:** Publicly available stock market data, including historical price, volume, and company fundamentals.
  - **Analysis/Modeling Ideas:** Create predictive models to forecast stock prices. Consider using time series analysis, sentiment analysis of news articles, or machine learning models like recurrent neural networks (RNNs) or support vector machines (SVMs).
4. **Healthcare Disease Diagnosis:**
  - **Data Source:** Publicly available medical records, patient data, and disease registries.
  - **Analysis/Modeling Ideas:** Develop machine learning models for disease diagnosis or risk prediction based on patient history, lab results, and demographics. Use techniques such as logistic regression, decision trees, or deep learning with CNNs or RNNs.
5. **Social Media Sentiment Analysis:**
  - **Data Source:** Publicly available social media data (e.g., Twitter, Reddit) related to a specific topic, brand, or product.
  - **Analysis/Modeling Ideas:** Analyze sentiment trends and public opinion. Use natural language processing (NLP) techniques to perform sentiment analysis, topic modeling, or sentiment-based recommendation systems.
6. **Energy Consumption Optimization:**
  - **Data Source:** Publicly available data on energy consumption patterns, weather data, and building information.
  - **Analysis/Modeling Ideas:** Create predictive models to optimize energy consumption in buildings. Use regression analysis to correlate energy use with external factors like weather and occupancy, or employ reinforcement learning for dynamic energy management.
7. **E-commerce Product Recommendation:**



	<ul style="list-style-type: none"> <li>• <b>Data Source:</b> Publicly available e-commerce transaction and user behavior data.</li> <li>• <b>Analysis/Modeling Ideas:</b> Develop recommendation systems using collaborative filtering, content-based filtering, or hybrid methods. Personalize product recommendations for users based on their past behavior and preferences.</li> </ul>
8.	<b>Traffic Congestion Prediction:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Publicly available traffic sensor data, GPS data, or traffic camera feeds.</li> <li>• <b>Analysis/Modeling Ideas:</b> Build models to predict traffic congestion patterns, helping commuters plan better routes and city planners optimize traffic flow. Use time series analysis, GIS, or machine learning models.</li> </ul>
9.	<b>Environmental Pollution Monitoring:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Publicly available environmental data, including air quality, water quality, and climate data.</li> <li>• <b>Analysis/Modeling Ideas:</b> Create models to predict pollution levels or analyze the impact of environmental factors on public health. Use regression analysis, time series analysis, or spatial analysis.</li> </ul>
10.	<b>Movie Box Office Prediction:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Publicly available movie industry data, including movie budgets, release dates, genre information, and historical box office performance.</li> <li>• <b>Analysis/Modeling Ideas:</b> Develop predictive models to estimate a movie's potential box office success. Explore factors like marketing budget, release timing, genre, and critical reviews. Use regression analysis or machine learning algorithms.</li> </ul>

Remember to choose a project that aligns with your interests and the availability of suitable data sources. The success of your data mining project often depends on the quality of data and the relevance of the analysis to the business question.

Can you re-generate the ideas, with links to specific datasets?

#### ChatGPT

I can provide you with ideas for data mining projects along with general categories of data sources, but I can't provide direct links to specific datasets as my knowledge is not up to date with the latest publicly available datasets. However, I can guide you on where to search for these datasets:

1.	<b>Customer Churn Prediction for a Telecom Company:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Look for customer churn datasets on platforms like Kaggle, UCI Machine Learning Repository, or explore telecom companies' open data initiatives.</li> </ul>
2.	<b>Retail Store Sales Forecasting:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> You can find retail sales and inventory datasets on Kaggle, government data portals, or open datasets from retail organizations.</li> </ul>
3.	<b>Stock Market Price Prediction:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Historical stock market data can often be found on financial data providers' websites, Yahoo Finance, or stock market APIs.</li> </ul>
4.	<b>Healthcare Disease Diagnosis:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Search for healthcare datasets on the UCI Machine Learning Repository, government health data portals, or research institutions' data repositories.</li> </ul>
5.	<b>Social Media Sentiment Analysis:</b>

	<ul style="list-style-type: none"> <li>• <b>Data Source:</b> For social media data, consider using APIs from platforms like Twitter, Reddit, or explore publicly available datasets on sentiment analysis on platforms like Kaggle or GitHub.</li> </ul>
6.	<b>Energy Consumption Optimization:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Some utility companies provide energy consumption data. Additionally, government agencies often publish energy consumption and weather data on their websites.</li> </ul>
7.	<b>E-commerce Product Recommendation:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Many e-commerce platforms offer APIs for accessing transaction and user behavior data. Alternatively, look for publicly available e-commerce datasets on Kaggle or research data repositories.</li> </ul>
8.	<b>Traffic Congestion Prediction:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Transportation departments in many cities provide traffic sensor data. You can also explore sources like traffic camera feeds or GPS data providers.</li> </ul>
9.	<b>Environmental Pollution Monitoring:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Government environmental agencies often publish air quality, water quality, and climate data. Look for environmental data portals and open data initiatives.</li> </ul>
10.	<b>Movie Box Office Prediction:</b> <ul style="list-style-type: none"> <li>• <b>Data Source:</b> Box office data is often available from movie industry databases or websites. You can also find movie-related datasets on platforms like Kaggle.</li> </ul>

When searching for datasets, be sure to check the data's licensing and usage terms to ensure it can be used for your project. Additionally, consider the size and quality of the dataset in relation to the complexity of your analysis or modeling tasks.

We're trying to answer the following business question using a TMDB dataset containing information such as budget, revenue, cast, director, etc. that are available through the API: 1. How should a production company decide to distribute a film, based on information available prior to release (and/or potentially initial reception?) (i.e. in theaters or VOD?) Which modeling techniques might we use to answer this question?

### ChatGPT

To answer the business question of how a production company should decide to distribute a film based on information available prior to release (and potentially initial reception), you can employ various modeling techniques and data analysis methods. Here are some modeling techniques and steps you can consider:

1. **Data Preprocessing:**
  - Data Cleaning: Clean the dataset by handling missing values, outliers, and inconsistent data.
  - Feature Engineering: Create relevant features such as genre, star power (based on cast), seasonality, and marketing budget.
2. **Exploratory Data Analysis (EDA):**
  - Perform EDA to understand the distribution of key variables like budget, revenue, and genre.
  - Analyze how different features (e.g., director, cast, genre) correlate with the film's performance.
  - Visualize trends and patterns that may provide insights into distribution strategies.
3. **Feature Selection:**
  - Use feature selection techniques to identify the most influential features that affect a film's distribution strategy.
4. **Predictive Modeling Techniques:**
  - a. **Classification Models:**
    - Binary classification models can be used to predict whether a film should be distributed in theaters or through Video on Demand (VOD).
    - Features can include budget, genre, cast, director, marketing budget, and release date.
    - Common classification algorithms include logistic regression, decision trees, random forests, and gradient boosting.
  - b. **Regression Models:**
    - If you want to predict a numerical value, such as expected revenue, you can use regression models.
    - Features can include budget, marketing budget, and other relevant factors.
    - Common regression algorithms include linear regression, decision tree regression, and random forest regression.
  - c. **Time Series Analysis:**
    - If you want to consider initial reception and track the film's performance over time, time series analysis can be useful.
    - Analyze revenue or audience reception trends to make decisions on continued distribution strategies.
5. **Cross-Validation:**
  - Use cross-validation techniques to assess the model's performance and ensure it generalizes well to new data.
6. **Model Evaluation:**
  - Choose appropriate evaluation metrics such as accuracy, precision, recall, F1-score, or Mean Absolute Error (MAE) depending on the type of model used.

	<ul style="list-style-type: none"><li>• Validate the model's performance on a holdout dataset or through k-fold cross-validation.</li></ul>
7.	<b>Hyperparameter Tuning:</b>
	<ul style="list-style-type: none"><li>• Optimize model hyperparameters to improve predictive performance.</li></ul>
8.	<b>Interpretability:</b>
	<ul style="list-style-type: none"><li>• Interpret the model to understand which features are most influential in determining the distribution strategy.</li></ul>
9.	<b>Deployment:</b>
	<ul style="list-style-type: none"><li>• Once you have a satisfactory model, deploy it for real-time or batch predictions to assist the production company in making distribution decisions.</li></ul>
10.	<b>Monitoring and Updates:</b>
	<ul style="list-style-type: none"><li>• Continuously monitor the model's performance and update it as new data becomes available.</li></ul>

Remember that the choice of modeling techniques and features should align with the specific goals and characteristics of your dataset. Additionally, it's essential to consider domain knowledge and industry expertise when making distribution decisions for films, as modeling is just one part of the decision-making process.

We're having some trouble envisioning how to classify whether a movie should be released in theaters or not, since we don't have information about how much money a movie made on a streaming platform (we do have it for the box office, however). We also know that obviously, a movie's success in the theater is linked to higher streaming numbers and so forth.

### ChatGPT

If you don't have direct information about how much money a movie made on a streaming platform, and you recognize that a movie's success in theaters is linked to higher streaming numbers and vice versa, you can approach the problem differently. Instead of directly classifying whether a movie should be released in theaters or not, you can build a predictive model that estimates the overall success of a movie based on available features (like budget, cast, director, marketing, etc.), and then use this estimate to inform distribution decisions.

Here's a revised approach:

1. **Define Success Metric:** Determine a success metric that represents the overall performance of a movie. This could be total revenue, profitability, or a composite score that takes into account both box office revenue and expected streaming revenue.
2. **Data Preparation:**
  - Clean and preprocess your dataset as previously mentioned.
  - Calculate or estimate the success metric for the movies in your dataset. For box office revenue, you already have the data. For streaming revenue, you may need to make assumptions or use proxies based on industry data or similar movies.
3. **Predictive Modeling:**
  - Use regression techniques to predict the success metric based on available features.
  - Features can include budget, genre, cast, director, marketing budget, and any other relevant factors.
4. **Threshold-Based Decision:** Set a threshold value for the success metric. This threshold will help you decide whether a movie should be released in theaters or not.
  - For example, if the estimated success metric is above the threshold, recommend a theatrical release; otherwise, recommend a streaming platform release.
5. **Validation and Evaluation:**
  - Validate the model's performance using historical data. You can use metrics like Mean Absolute Error (MAE) or Root Mean Square Error (RMSE) to assess how well the model predicts the success metric.
6. **Decision Support:**
  - Use the trained model to make distribution recommendations for new movies. When a new movie is in production, input its features into the model to estimate its expected success metric.
  - If the estimated success metric exceeds the threshold, suggest a theatrical release. If not, recommend a streaming platform release.
7. **Monitoring and Updates:**
  - Continuously monitor the model's performance and update it as new data becomes available.

This approach allows you to make distribution decisions based on the expected overall success of a movie, considering both box office revenue and estimated streaming revenue, even if you don't have direct streaming revenue data. It recognizes the interconnectedness of these two revenue streams and provides a data-driven framework for distribution decisions.