ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/kailash-alle/198

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 1

Research Article

Predictive Analytics for Box Office Success

Kailash Alle

Information Technology (IT) / Software Development, Sr. Software Engineer, Comscore Inc, USA

Citation: Alle K. Predictive Analytics for Box Office Success. *J Artif Intell Mach Learn & Data Sci 2022*, 1(1), 800-804. DOI: doi.org/10.51219/JAIMLD/kailash-alle/198

Received: 02 June, 2022; **Accepted:** 18 June, 2022; **Published:** 20 June, 2022

***Corresponding author:** Kailash Alle, Information Technology (IT) / Software Development, Sr. Software Engineer, Comscore Inc, USA, Tel: 626-693-7845; E-mail: kalle@comscore.com

ABSTRACT

We aim to predict a movie's profitability to help with investment decisions early in the production process. By using data from different sources and applying social network analysis and text mining techniques, our system looks at various features such as who is in the cast, what the movie is about, and when it will be released. Our experiments show that our system performs much better than standard methods. The new features we introduced significantly improved our predictions. Besides creating a useful decision support system, we also studied the main factors that affect a movie's profitability. Additionally, we demonstrated how our system can recommend cast members to maximize profits. This research shows the power of data analytics in helping businesses make better decisions.

Keywords: Predictive Analytics, Box Office

1. Introduction

Movies play a major role in our lives, serving as a key medium for conveying messages, scientific innovations, stories, history, culture, and entertainment. Given their significance and widespread popularity, knowledge and research about the film industry are growing rapidly. Each year, hundreds of movies are released, ranging from low-budget to high-budget productions. Some become blockbusters, while others end up as flops or receive average ratings, depending on their budget, box office performance, and user reviews.

High-budget movies often involve multiple investors, and producers must persuade them to invest in projects that could either be huge successes or major failures-an unpredictable decision during the early stages of production. Predicting a movie's success has been extensively studied, with various useful datasets available to gauge a film's potential. Our research focuses on predicting a movie's success or failure specifically in terms of capital investment and the revenue it will generate.

2. Movie Background

The motion picture industry is a multibillion-dollar business, with the United States and Canada earning over $11.1 billion in box office revenues in 2015. Despite this, the financial success of a movie is highly unpredictable, with both hits and flops released each year. Researchers have tried to predict movie success using various methods, often focusing on box office revenues or theater admissions. However, investors want assurance that their investments will yield returns. For example, while "Evan Almighty" earned $100 million, it cost $175 million to produce. In contrast, "Super Troopers" cost $3 million but earned $18.5 million, making it a better investment. Between 2000 and 2010, only 36% of U.S. movies had box office revenues higher than their production budgets, highlighting the need for better investment decisions.

Our research defines a movie's success as its profitability and aims to predict this success in an automated way to support investors. The movie production process begins with the development phase (script and screenplay creation), moves to

preproduction (team assembly, location scouting, and investment securing), then to actual production (filming), postproduction (editing and effects), and finally distribution. For investment decisions, profitability predictions must be made before the production phase, using data available at that time.

This research proposes a Movie Investor Assurance System (MIAS) to predict movie profitability early in the preproduction phase. Using historical data, the system extracts key characteristics such as who is involved, what the movie is about, and when it will be released. It then uses machine learning to predict success based on profitability criteria. Our study focuses on using preproduction data to make these predictions, as predictions made later in the process, although more accurate, are too late for investors.

We present the first system to predict movie profitability at an early stage, demonstrating how various types of data can be used to train machine learning algorithms. This approach can provide powerful forecasts and recommendations for business decisions. Additionally, we introduce new features like dynamic network features, plot topic distributions, and profit-based star power measures, which significantly enhance the system's performance and help explain factors behind a movie's profitability.

3. Defining Success

Defining what makes a movie successful is crucial to this issue. Previous studies have mostly focused on gross box office revenue[3,4,18,28,30,34] or the number of admissions[5,26] as metrics of success. The idea is simple: if a movie sells well at the box office or attracts a large audience, it's considered successful. However, these metrics overlook the costs involved in making the movie. Our analysis of historical data confirms that high revenues do not always translate to high profits (more details will be discussed later). Therefore, a more meaningful measure of success should consider profitability-either as the actual profit earned[36] or the return on investment (ROI)[14].

Once a success metric is chosen, studies often categorize movies into successful or not successful based on revenues, treating this as a binary classification task. Some studies have approached it as a multiclass classification problem, categorizing movies into different success categories [30]. Additionally, predictions have been made using continuous numerical values of success metrics[17,28,38], with some studies using logarithms of these metrics to handle their distribution[34,36,40].

4. Factors Affecting Movie Success

The accuracy of a predictive model depends heavily on the features it uses-these are the independent variables that influence predictions. In studying what makes a movie successful, researchers have explored three main types of features: audience-based, release-based, and movie-based features.

4.1. Audience-based features

Focus on how potential viewers perceive a movie. Positive reactions from audiences, whether on platforms like Twitter, trailer comments, blogs, news articles, or movie reviews, often indicate higher box office revenues.

4.2. Release-based features

Consider when and how widely a movie is released. Factors like the number of theaters showing the movie and the timing of its release-such as during holidays or specific seasons-can affect its financial success. Competition during the release period also plays a role.

4.3. Movie-based features

Are characteristics directly related to the movie itself. This includes who stars in it and its storyline. For instance, movies featuring well-known actors often perform better financially. However, our research suggests that an actor's profitability record might be more crucial than just their popularity in predicting a movie's financial success.

Directors also play a significant role, though they are sometimes overlooked compared to actors. Studies show that both the star power of actors and the collaboration dynamics among the cast can influence a movie's profitability. Our research expands on this by examining how director involvement affects financial success.

In addition to individual factors, we explore how different types of features interact. For example, how the combination of actors' popularity and their experience in various genres influences a movie's success. We also consider broader market trends and genre popularity over time.

Our approach integrates traditional features like actors' earnings and new ones such as team expertise and diversity. We use advanced techniques like text mining and social network analysis to automatically extract and analyze these features from large datasets. This allows us to predict a movie's profitability early in its production cycle, leveraging data available at that stage.

Furthermore, our study challenges previous findings by examining whether factors like actors' and directors' star power and teamwork dynamics hold true when profitability, rather than just revenues, is the measure of success. This is based on a comprehensive analysis of a much larger dataset, aiming to provide more accurate insights for decision-making in the film industry.

5. Data Preparation

We primarily used the IMDB Movies Extensive Dataset, which covers movies from 1894 to 2020 with at least 100 votes on IMDb. IMDb holds information on over 6 million titles, including nearly 500,000 featured films. The dataset consists of five CSV files:

1. **Movies.csv**: Contains details on 85,855 movies with 22 attributes.
2. **Names.csv**: Lists 297,705 cast members with 20 attributes.
3. **Ratings.csv**: Provides rating details for 85,855 movies with 49 attributes.
4. **Title_principals.csv**: Includes 835,513 casting roles with 6 attributes.
5. **Awards.csv**: Contains 1,885,525 records with 21 attributes related to awards.

These datasets were merged into one comprehensive dataset for our research. In addition to the existing columns, we created derived columns to enhance our model's accuracy. These include count actors, actor_1_name, actor_2_name, actor_3_name, count directors, director name, count producers, producer name, count writers, and writer name. These variables were

chosen because we believe that a strong cast and skilled director significantly benefit a movie's success in business terms.

To standardize the dataset, we selected the top three actors and the top director, writer, and producer for each movie, based on the ordering column in the title principals dataset. In cases where there were fewer than three actors or multiple directors, writers, or producers, we adjusted accordingly.

Additionally, we utilized the awards dataset to derive a crucial attribute called "star power." This aggregate value reflects the achievements and accolades of actors, directors, writers, and producers, further enhancing our model's predictive accuracy. Attributes such as actor_1_wins, actor_1_nominations, actor_2_wins, actor_2_nominations, actor_3_wins, actor_3_nominations, director wins, director nominations, producer wins, producer nominations, writer wins, writer nominations, film wins, and film nominations were derived from this dataset to quantify the influence of these factors on movie profitability.

5.1. Data transformation and filtering

We focused on predicting the Box Office revenue or Worldwide Gross Income as our target variable. Initially, we standardized key monetary variables like budget and USA gross income to US Dollars using the currency converter Python library. To ensure fairness, we only included movies from 1990 onwards, avoiding bias from older data.

Since our study specifically targeted Bollywood and Hollywood movies, we filtered out films in languages other than Hindi and English. This filtering process left us with 28,163 rows and 59 columns. Not all columns proved useful; for instance, many movies lacked data beyond their primary genre (genre_1), leading us to drop genre_2 and genre_3 columns due to extensive null values. Similarly, columns with more than 80% null values were removed.

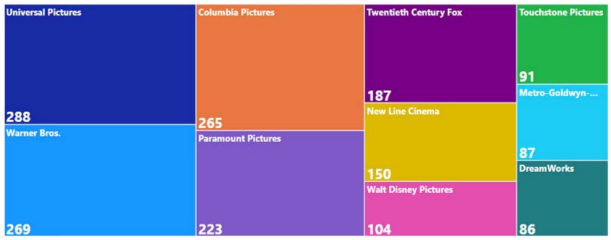
Following this cleanup, we numerically imputed missing values for variables like budget, average votes, and meta score based on the movie's release year. We then analyzed correlations between these variables and revenue, removing columns with correlation coefficients between -0.2 and +0.2. Finally, our dataset was refined to 28,150 rows and 39 columns for further analysis.

5.2. Exploratory data analysis

To explore our data, we used the Plotly library in Python and Microsoft Power BI. Our analysis focused on several aspects: trends in budget and box office revenue across different paradigms such as production companies, year-over-year growth or decline, star power impact, genres, and IMDb ratings.

Figure 1 illustrates the distribution of movies among various production houses from 1990 to 2020. It shows how these houses share the movie market.

Figure 2 presents a pie chart detailing box office revenue and budget distribution among top production houses. Notably, Marvel Studios stands out with fewer movies but significantly higher revenue compared to budget.

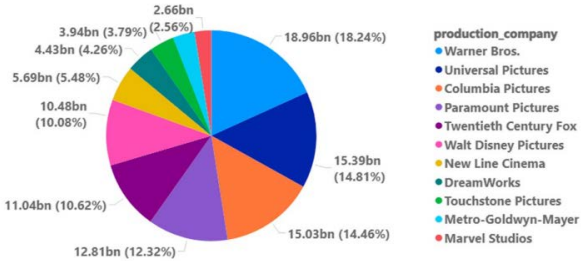


The total budget spent is approximately $2.66 billion, whereas the worldwide gross revenue amounts to $15.06 billion.

Figure 3 breaks down movies released from 1990 to 2020 by genre. "Drama" emerges as the most popular genre, capturing about 26.3% of the market, followed by Comedy and Action genres. "Fantasy" holds the smallest share at 3.32%.

Figure 4 tracks year-on-year trends in production house performance, revealing Marvel's consistent revenue growth, especially post-2010s. Conversely, Warner Bros. and Universal Pictures show declines, possibly due to competitive pressures or varying movie performances.

Examining **Figure 5** also shows a significant rise in movie profitability from the early 2000s to mid-2019, followed by a downturn in late 2019, likely due to COVID-19's impact on cinema operations and movie releases moving to online platforms.

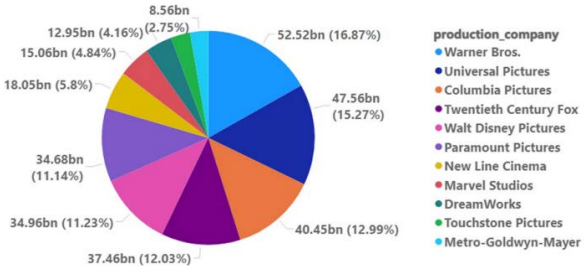


5.3. Budgets utilized by top production houses

Depicts the yearly distribution of budgets and revenue generated over 30 years, highlighting a noticeable drop in 2020 due to COVID-19.

Next, we explored the relationship between budget and box office revenue, revealing a linear correlation where higher budgets tend to correlate with higher revenues.

Similarly, Figure 6 explores the relationship between revenue and IMDb ratings, showing another linear correlation where higher-rated movies often yield higher revenues.



5.4. Revenues generated by top production houses

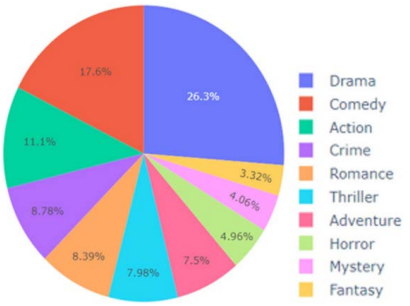
Contrarily, exploring whether highly-rated movies necessarily earn high revenues (Figure 6) reveals nuances. For instance, "The Shawshank Redemption," with an IMDb rating of 9.3, earned only $28 million despite a $25 million budget, indicating that high ratings don't always guarantee high box office earnings.

5.5. YoY trend of revenue generated by production companies

Next, we want to explore how the box office revenue relates to other aspects of a movie, like the director, writer, and main actor. The following graphs show the impact of these roles on movie profitability.

Figure 7 presents statistics on the top 10 directors who have

directed commercially successful films. James Cameron stands out, having directed blockbuster movies like Titanic and Avatar, which earned significant revenue with relatively low budgets, despite directing only four films since 1990.



Genre wise split

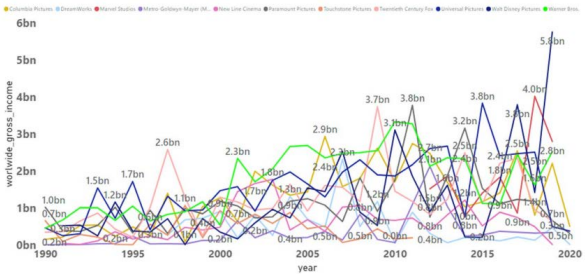


Figure 8 examines the influence of writers on movie profitability. Christopher Marcus, known for writing just nine movies, has contributed scripts that collectively generated over $8 billion in revenue.

Similarly, Figure 9 explores the impact of primary actors on movie profitability. Robert Downey Jr., primarily known for his role in Marvel movies, has been a major revenue generator. This illustrates that famous actors can significantly boost a movie's financial success, regardless of the number of films they've starred in.

In conclusion, these examples highlight a strong correlation between these roles and a movie's potential success or failure, as well as its revenue performance.

5.7. Data conversion

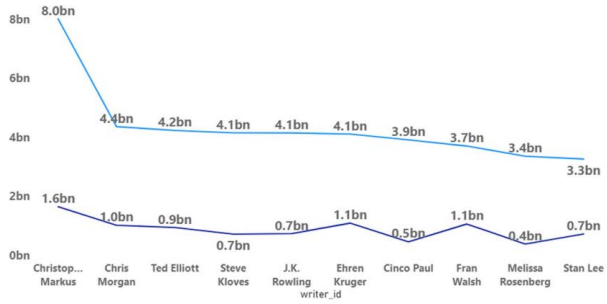
The success of the predictive model doesn't just rely on the algorithm and settings but also on how the data is structured. Better data leads to better results, while poorly formatted data can hinder performance. Typically, machine learning models require numerical data. Since our dataset includes continuous variables such as actor names, director names, and producer names, we used encoding methods to convert these features into numerical formats. This ensures that the models can process the data correctly.

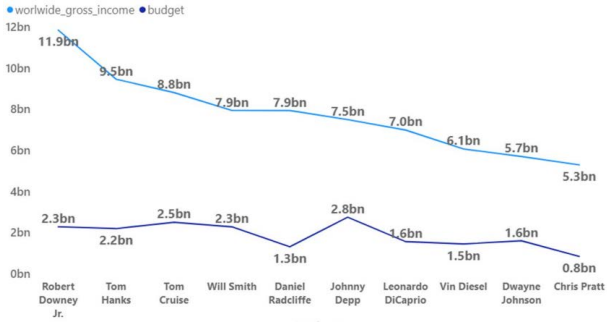
5.8. Label Encoding

We used Label Encoding to convert continuous features into numerical format. This technique is suitable when there are inherent categories in the dataset. For instance, actor names act as categories because we're studying the impact of specific actors in movies. Label Encoding ensures that each actor is assigned a consistent integer value. This allows our model to accurately predict how casting the same actor in different movies affects their performance at the box office.

5.9. Performance metrics

Evaluation metrics help us assess how well the proposed model performs. In this experiment, where we used several algorithms and even employed ensemble learning, we used the following metrics to evaluate the model: R2 score, Mean Absolute Error, Mean Squared Error, Explained Variance Score, Root Mean Squared Error, and Normalized Mean Squared Error.





- R^2 Score:

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}.$$

- Mean Absolute Error:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

- MSE:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

- Explained Variance Score:

$$= 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

- RMSE:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

5.10. Trial

Here are the results from each technique we used, where we allocated 30% of the data for testing and 70% for training.

6. Linear Regression

Linear regression uses training data points to find the best-fit line that minimizes residual error, which is the sum of differences between predicted and actual values.

In supervised learning, like linear regression, the dataset includes both independent (input) and dependent (output) columns. To predict the output accurately, the model must be trained with relevant input columns. Selecting effective features is crucial for model performance. Columns highly correlated with the output variable are chosen, while irrelevant ones are omitted to avoid harming the model's accuracy. You can determine correlation by calculating the correlation coefficient

between each column and the output variable. A coefficient close to +1 or -1 indicates strong correlation, making those variables important for the model. A coefficient near 0 suggests little dependency.

The equation of a linear regression line is Y = a + bX, where X is the independent variable and Y is the dependent variable. Here, b represents the slope of the line, and a is the intercept (the value of Y when X equals 0).

In our problem, the algorithm provides these results using 30% of the data for testing and 70% for training:

Model Score % (R2 * 100)=69.76% R2 Score=0.697569 Mean Absolute Error (MAE)=2.497535e+07 Mean Squared Error (MSE)=2.181325e+15 Explained Variance Score (EVS)=0.697571 Root Mean Square Error (RMSE)=4997.534699 Normalized MSE=2960.568168 Max Error=1.020606e+09 Mean Absolute Percentage Error=842.813880

7. Stacking Regression

Stacked generalization, also known as stacking, is an ensemble technique designed to improve model performance by combining the predictions of multiple models.

Here's how it works:

- **Base Models (Level-0 Models)**: These are individual models trained on the training data, each making its own predictions.
- **Meta-Model (Level-1 Model)**: This model learns how to best combine the predictions from the base models to produce a final prediction.

The meta-model is trained using the predictions made by the base models on data that they did not see during training. The inputs to the meta-model are typically the predictions made by the base models, which can be real values (for regression tasks) or probabilities/class labels (for classification tasks).

In our study, we used several base models: LGBMRegressor, XGBRegressor, CatBoostRegressor, and GradientBoostingRegressor. These models generate predictions independently. The meta-model, in our case, is LinearRegression, which takes the predictions from the base models and combines them to make a final prediction.

For our experiment, the algorithm provided these results using 30% of the data for testing and 70% for training:

8. Conclusion

Predicting the success of movies is crucial because it impacts not only the actors but also the production companies and producers involved. Many studies in this field have used different machine learning techniques, focusing on traditional features, social media reactions, critical reviews, and more. In this research, we took a diverse approach, incorporating both traditional and derived variables such as the influence of actors, producers, and directors, along with their awards and nominations. We discovered that specific factors, like casting a particular actor, significantly influence a movie's revenue.

To validate our findings, we employed multiple machine learning algorithms and evaluated them using various metrics. Among these, CatBoostRegression and Stacking Regression stood out, achieving the highest model accuracies of 83.84% and 83.5%, respectively. Since CatBoostRegression is an ensemble model itself and performed better than stacking regression, it suggests that using inherent ensemble models yields superior results, eliminating the need to manually stack individual models to create another ensemble.

9. Limitations

This research is limited to using traditional features and their historical and derived aspects, such as counting actors, awards received by the lead actor, producer, director, etc. However, future studies could expand by integrating social media factors, as seen in[1,2,6,8,12,14,20]. Our initial goal was to predict revenue accurately using regression techniques, focusing solely on these traditional features and their derivatives.

In the future, incorporating factors like fan followings of key figures, YouTube likes or comments on previous movies involving these figures, or even sentiment analysis of critic and public responses could significantly enhance predictive accuracy. While this scope was beyond our current research, it holds potential for future implementations.

10. References

1. Abbasi MA, Memon ZA, Durrani NM et al. A multi-layer trust-based middleware framework for handling interoperability issues in heterogeneous IOTs. Cluster Comput 2021;24: 2133-2160.
2. Bhave A, Kulkarni H, Biramane V, Komsakar P. Role of diferent factors in predicting movie success. International Conference on Pervasive Computing 2015.
3. Bistri WR, Zaman Z, Sultana N. Predicting IMDb rating of movies by machine learning techniques. International Conference on Computing, Communication and Networking Technologies 2019.
4. Bosse T, Memon ZA, Treur J. Emergent storylines based on autonomous characters with mindreading capabilities, 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'07), IEEE 2007; 207-214.
5. Bosse T, Memon ZA, Treur J, Umair M. An adaptive human-aware software agent supporting attention-demanding tasks. In: Yang J-J, Yokoo M, Ito T, Jin Z, Scerri P (edn.), Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems. Springer 2009;5925: 292-307.