Predicting Box Office Returns for Movies Based on Metadata and Electronic Word-of-Mouth Metrics

Casey Derringer
The University of Alabama
Tuscaloosa, AL
cjderringer@crimson.ua.edu

Samuel Wisnoff
The University of Alabama
Tuscaloosa, AL
slwisnoff@crimson.ua.edu

Garrett Weakley
The University of Alabama
Tuscaloosa, AL
gsweakley1@crimson.ua.edu

ABSTRACT
The film industry is a complex ecosystem in which many creative, technical, and market factors influence a movie's financial success. Studios routinely invest significant resources into forecasting box office performance, yet traditional intuition based on heuristic approaches struggle to capture the nonlinear relationships among production attributes, audience preferences, and early reception signals. This paper focuses on the domain of data driven predictive modeling for cinema, addressing the problem of accurately estimating box office revenue before a film is released.

To tackle this challenge, we leverage a large-scale dataset containing metadata for hundreds of thousands of movies, including production details, cast and crew information, budgetary variables, and genre classifications. This dataset is supplemented with electronic word-of-mouth (EWOM) indicators collected from Youtube, enabling us to incorporate real time audience interest signals into the modeling process. Together, these features form a rich, multimodal input space suitable for regression-based revenue prediction.

Our solution approach centers on constructing and evaluating machine learning models capable of mapping this feature space to realistic revenue estimates. We preprocess and encode metadata, integrate EWOM features, and compare several algorithms to identify the most effective architecture. Model training emphasizes a balance between prediction performance and robustness to noise, meant to tackle the inherently unpredictable nature of film performance.

The implemented system allows users to input relevant metadata for an upcoming film and receive a projection of box office returns. This tool is designed to support filmmakers, producers, and studio executives in making data driven decisions throughout the development and marketing process.

Evaluation results demonstrate that incorporating both structured metadata and EWOM signals produces strong proofs-of-concept for predictive models, suggesting that data driven approaches can meaningfully enhance forecasting reliability in an industry defined by uncertainty.

Author Keywords
Box office prediction, Movie analytics, Decision trees, Data mining, YouTube engagement

INTRODUCTION
The movie industry has long functioned as both a cultural cornerstone and major economic force, contributing greatly to domestic and international markets. In recent years, however, the sector has faced substantial volatility. Domestic box office revenue has declined by nearly thirty percent since the onset of the COVID-19 pandemic [1], mirroring similar downturns across global markets. This contraction has heightened the financial uncertainty surrounding film production, where studios invest immense resources despite limited guarantees of commercial success. Traditional forecasting approaches grounded in executive intuition, historical comparisons, or test screening responses have become increasingly inadequate in an environment shaped by shifting consumer behavior, intensified competition from streaming platforms, and rapid changes in promotional ecosystems. As a result, films frequently underperform relative to expectations, causing financial losses and public scrutiny.

The central challenge for industry stakeholders is determining which factors meaningfully influence box office outcomes and how those factors interact. Prior research has addressed components of this problem, including efforts to model country level box office markets [2] and studies predicting film revenue using relatively modest datasets [3]. While informative, these approaches often rely on narrow feature sets or omit modern engagement signals, limiting their applicability to contemporary filmmaking practices. Moreover, existing studio workflows rarely integrate diverse data sources such as production metadata, cast and crew histories, release timing, and electronic word-of-mouth (EWOM) indicators into a unified predictive framework. This gap underscores the need for a scalable, data driven system capable of capturing the complex and nonlinear determinants of box office performance.

This project proposes such a system by combining large scale movie metadata with EWOM metrics to generate more reliable revenue predictions. The approach leverages the Ultimate 1 Million Movies Dataset in conjunction with YouTube trailer engagement data collected through an Apify web scraper. Together, these sources create a multimodal feature space that incorporates both structured production

information and dynamic measures of audience interest. By integrating these components, the system seeks to provide more accurate and interpretable forecasts than traditional heuristic based methods.

We used scikit-learn to build a box office revenue predictor, training a model on data scraped from youtube on excitement surrounding the movie's trailer and a dataset from Kaggle pulling off TheMovieDB's website. We then deployed this model as a dashboard with gradio, providing an intuitive way for a user to enter all the necessary data for the model and receive their prediction for target revenue. Our model was able to achieve a near 0.7 R^2 score on a 0-1 scale, along with a Mean Absolute Error of approximately 25 million, offering a stable revenue prediction when frequently dealing with targets in the hundreds of millions and sometimes billions.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on box office forecasting, EWOM analytics, and machine learning applications in film studies. Section 3 describes the architecture of the predictive system, including data preprocessing, feature engineering, and model design. Section 4 presents the evaluation methodology and empirical results.

RELATED WORK

Other papers in this area have slightly different targets and datasets. [2] uses an SVM-based method to predict the overall box office return of entire markets, such as the U.S. and China, instead of individual movies, requiring much different data. [3] examines a Random Forest approach to predicting box office returns on specifically chinese movies from 2017 to 2019, analyzing a much smaller set of data on a different domestic market altogether.

Prior research on box office forecasting spans multiple methodological approaches and problem scopes, offering valuable insight into both the challenges and opportunities for data driven prediction systems. Much of the existing literature highlights the influence of electronic word-of-mouth (EWOM), online engagement behavior, and metadata signals on audience decisions, yet important gaps remain in dataset scale, feature diversity, and generalizability. A comparative examination of these works underscores the need for a more comprehensive, multimodal system such as the one proposed in this study.

Several studies emphasize the importance of EWOM in shaping consumer demand. [5] examines broad trends in digital social interaction and its effects on box office outcomes, identifying consistent positive correlations between online discourse and financial performance. This work demonstrates the significance of user generated content for understanding market reception, but it focuses on conceptual trends rather than predictive modeling. In contrast, [11] evaluates how the helpfulness of online reviews moderates the relationship between EWOM and box office performance. Their findings suggest that not all EWOM signals contribute equally to predictive accuracy, underscoring the importance of data quality. Building on these insights, our system integrates high signal EWOM metrics to capture early engagement patterns mean to reflect audience interest.

A second cluster of related research investigates machine learning models trained on either market level or country specific data. [6] employs an SVM-based method to forecast overall national box office revenues in the United States and China. Although their model demonstrates strong performance for large scale, macroeconomic prediction, it does not address the more granular challenge of estimating revenue for individual films. Similarly, [12] applies supervised learning methods to classify movie popularity, offering insight into metadata driven classification tasks but not regression forecasting. These works reveal the value of classical ML techniques, yet they do not utilize multimodal feature sets or the scale of data available in contemporary film datasets.

Other researchers have explored movie level prediction using more targeted datasets. [7] incorporates online reviews and web search trends to estimate film performance, showing that models combining multiple sources of audience information outperform those relying on a single modality. [10] expands on this idea by integrating social network signals and topic modeling, demonstrating that deep learning models can effectively merge structured features with review based text. [10] further investigates neural network architectures, combining emotional mining with metadata to improve prediction accuracy. While these studies showcase the strengths of multimodal feature integration, their datasets are generally limited to narrow time windows, specific national markets, or relatively small samples.

The system proposed in this work builds on the strengths of prior research while addressing their limitations. Unlike market level studies [6] or narrow domestic analyses [10], our system draws from the extensive, globally representative Ultimate 1 Million Movies Dataset, supplemented with high coverage EWOM engagement metrics. Compared with neural and ensemble approaches trained on smaller datasets [9], our pipeline applies scalable preprocessing, metadata encoding, and regression modeling to a significantly larger corpus. This enables broader generalizability across genres, markets, and release periods.

Collectively, the literature underscores both the promise and the limitations of existing forecasting systems. Prior studies demonstrate that metadata, EWOM, and machine learning each contribute meaningfully to understanding box office performance, yet no single study integrates all three at scale. By combining large scale metadata with contemporary EWOM indicators in a unified prediction pipeline, our project fills this methodological gap and advances toward a more comprehensive and operationally useful forecasting tool for industry stakeholders.

SYSTEM ARCHITECTURE

The following subsections describe the design vision and goals for the system architecture of the project.

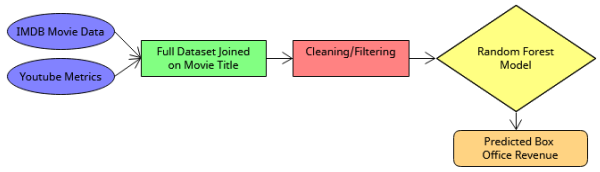

Figure 1. System architecture showing the steps from collection of data through a predicted Revenue from the model.

Design Vision

The implemented system solves the real-world problem of ensuring the commercial success of movies; large amounts of time and money are invested into movie production with the goal of commercial success. Failure to meet revenue expectations may lead to disastrous results for cast, crew, and investors. The design goal of the model is to predict how electronic word of mouth will impact the revenue.

Datasets

Two primary datasets formed the basis of this project. The first was "The Ultimate 1 Million Movies Dataset" available on Kaggle available for use under the Apache 2.0 license [5]. The second dataset used was obtained using the "YouTube Scraper" from the website Apify. This dataset consists of data scraped from YouTube trailers of around 7000 movies that were also present in the Kaggle dataset. These two datasets were consolidated to form a single working dataset that could be worked on as a single pandas dataframe.

Work had to be done to ensure that only data that would be relevant to a movie receiving box office revenue was scraped from the YouTube trailers. In this case, the model used the view and like count for each trailer. While there is other data that can be extracted from YouTube videos such as comment sentiment or video runtime, these were deemed costly for the prediction model to process and largely unimportant compared to the views and likes which capture a relatively accurate picture of how popular a certain movie may be, especially when considering whether people will go to see the movie and contribute to the box office revenue.

Candidate Feature Sets and Prediction Targets

While there was an abundance of movie data available given the datasets used, there were many movies that had to be cut to ensure that the prediction models would work efficiently with modern movies. The initial changes made to the dataset included: removing movies unreleased or released prior to 2012, movies with 0$ revenue, and movies without a title in the Kaggle dataset. These steps ensured the data was not influenced by out of date, or irrelevant information.

A large challenge with feeding movies into the model was the amount of data and data types available. Many features such as cast lists or genres had too many possible categories to be useful without encoding. For the genres, the solution was a one-hot encoding to store the genres as a binary "1" if the genre was present, and "0" if not. The solution for the other categories used a different method. For the cast, directors, writers, and production company, leave-one-out target encoding with M-estimate smoothing was used. This method represents the members that make up this category with the average of movie revenue for every movie that they are a part of in the dataset, excluding the current entry being computed. For composers, we represented them as a binary value if they were a "top" composer or not, and the director of photography was represented as an integer quantifying their experience in the dataset. This serves to remove the lengthy and ineffective process of sorting through large lists of names for each movies and instead allows the model to operate on numerical values for these categories.

Some other feature engineering included boolean features denoting whether the movie's primary language was English, it was a part of a pre-existing franchise, or if it was released during "peak season", generally referring to the summer and holiday months, all contributing to provide more valuable context to the model, and eliminate noise.

Learning Algorithms

The learning algorithms used in the final system were decision tree regression and random forest regression. Both models were implemented for comparison between methods. The inputs into the learning algorithms were the features such as budget, cast information, and YouTube metrics, among others. These algorithms were selected to predict the continuous target variable 'revenue' because of their non-linearity and strength in regression problems.

Decision Trees are a tree-based model that separates data recursively based on feature values to minimize the mean squared error loss function. The intermediate nodes of the tree have feature thresholds that determine how the observations are split amongst the trees [14]. The final leaf node where the input observation is sorted is the average value of all observations that were sorted to that leaf node by the decision tree [14]. The justification for applying decision trees is their ease of interpretability and ability to handle non-linear relationships through the intermediate node split process. The Decision Tree Regressor was implemented with a maximum depth of 5 with a minimum of 10 samples per leaf.

Furthermore, a Random Forest ensemble model was implemented to provide a more robust regressor than a singular decision tree. Random Forest regressors are composed of many decision trees that are each trained on subsets of the training data [15]. The decision trees trained are then averaged to form a final model more resilient to high variance features [15]. The justification for using Random Forest Regression is the potential for a more robust model. The Random Forest regressor was implemented with a maximum depth of 8, a minimum of 10 samples per leaf; the regressor is composed of 100 decision trees.

Both models intake all the numerical feature sets, including the encoded attributes for categorical data. For the target value of revenue, two metrics are used to evaluate the model performance: R^2 and Mean Absolute Error. The R^2 coefficient of determination is a score for the regression model with a maximum value of 1, indicating perfect performance, while Mean Absolute Error is the average distance between the predicted value and the actual value [16]. Both of these metrics provide insight into the performance of the Decision Tree and Random Forest regression models.

The software libraries used to implement both models are sklearn and statistics libraries. The package sklearn includes both the Decision Tree and Random Forest regressors. The library allowed customization of model parameters such as maximum depth and maximum leaves [17], and repeated k-fold cross validation verifying the performance of our features. The statistics package provided methods to calculate the metrics used to evaluate the model output. Both the average and median R^2 and Mean Absolute Error values were calculated for evaluation and comparison of models [18].

User Interface

Our User Interface takes the form of a Gradio app built into our data science pipeline. This affords the ability for a rapidly iterable and deployable interface for us test and share our model with the broader community. It uses a dual column setup, with fields for the user to enter all of the necessary features the model needs to produce a prediction. From there, the user can select the "Run" button below all the feature to generate a prediction from the model, then subsequently see the model's prediction displayed below that.

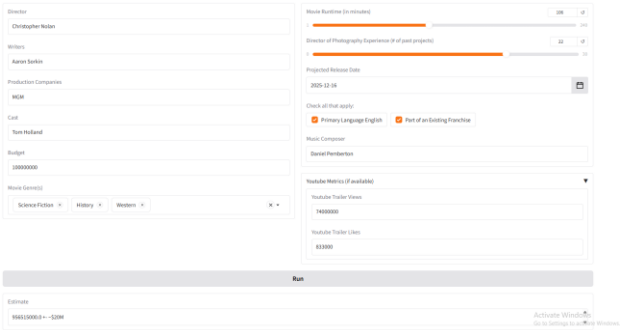

Figure 2. Gradio interface with dual column feature entry, a button for running the model, and a textbox to display predictions

SYSTEM EVALUATION

Metric/Model	Decision Tree w/out YouTube (YT) data	Decision Tree with YT data	**RF (w/out YT Data)**	**RF (with YT Data)**
MAE	$23855075	$30302554	$20554804	$24557612
R^2	0.4743	0.4579	0.6483	0.6947

Table 1. Evaluation outcomes. R^2 and Mean Absolute Error for Decision Tree and Random Forest regression with and without YouTube metrics.

When comparing the Decision Tree and Random Forest regression models, Random Forest consistently produced higher R^2 values and lower mean absolute errors, highlighting the robust nature of Random Forest compared to singular decision trees. With the addition of the YouTube data, the model improves slightly in terms of R^2 and mean absolute errors values, but only by about seven percent for Random Forest and by about two percent for Decision Trees in terms of R^2 values.

The report's initial expectations were that YouTube metric data would provide additional information to increase model performance. The results show that this increase in performance did occur, but at a smaller scale than expected, and with only certain models. Table 1 shows the comparison between the model output with and without YouTube data, revealing that a small boost in performance occurred with the addition of YouTube data on the Random Forest Model. Interestingly, the Decision Tree model saw a slight decrease in performance with the addition of the YouTube data, perhaps signaling underfitting.

For future work to improve the model, we will expand the different feature set to include aggregate critic scores of the movies. These scores can play a part in whether a potential audience member decides to see it, making them important learning features. Further tuning of hyperparameters may also yield higher performance metrics. In conclusion, the Decision Tree and Random Forest regressors gained a small performance boost with the addition of YouTube analytics on top of traditional movie features.

PROTOTYPE

Here is the Google Colab link to the working prototype of the final project:

https://colab.research.google.com/drive/18IcM6kZYJgfq83sleA-MdJzv07WK6tLK?usp=sharing.

REFERENCES

[1] Box Office Mojo by IMDbPro. Domestic Yearly Box Office. November 18, 2025. Retrieved November 18, 2025 from https://www.boxofficemojo.com/year/?ref_=bo_nb_di_secondarytab

[2] Li, D., & Liu, Z.-P. (2022). Predicting Box-Office Markets with Machine Learning Methods. Entropy, 24(5), 711. https://doi.org/10.3390/e24050711

[3] He, Qi, Hu, Bin, Research on the Influencing Factors of Film Consumption and Box Office Forecast in the Digital Era: Based on the Perspective of Machine Learning and Model Integration, Wireless Communications and Mobile Computing, 2021, 6094924, 10 pages, 2021. https://doi.org/10.1155/2021/6094924/

[4] Zhang, Z., Meng, Y. & Xiao, D. Prediction techniques of movie box office using neural networks and emotional mining. Sci Rep 14, 21209 (2024). https://doi.org/10.1038/s41598-024-72340-z

[5] Qiao, W. (2024). From Digital Social to Box Office Revenue: Analysis of the Impact and Trends of Electronic Word-of-Mouth. Galactica Media: Journal of Media Studies, 6(4), 426-453. https://doi.org/10.46539/gmd.v6i4.500

[6] Li D, Liu ZP. Predicting Box-Office Markets with Machine Learning Methods. Entropy (Basel). 2022 May 16;24(5):711. doi: 10.3390/e24050711. PMID: 35626594; PMCID: PMC9141781.

[7] Wang, X., & Li, R. (2022). Enhancing Film Box Office Predictions: Integrating Online Reviews and Web Search Trends. Journal of Computer Science and Software Applications, 2(1), 6–15. https://doi.org/10.5281/jcssa.v2i1.48

[8] Xie,C. (2024). A refined approach to early movie box office prediction leveraging ensemble learning and feature encoding. Applied and Computational Engineering,75,273-284.

[9] Zhang, Z., Meng, Y. & Xiao, D. Prediction techniques of movie box office using neural networks and emotional mining. Sci Rep 14, 21209 (2024). https://doi.org/10.1038/s41598-024-72340-z

[10] Chen Y, Dai Z. Mining of Movie Box Office and Movie Review Topics Using Social Network Big Data. Front Psychol. 2022 May 26;13:903380. doi: 10.3389/fpsyg.2022.903380. PMID: 35693503; PMCID: PMC9178289.

[11] Lee, S., Choeh, J.Y. The impact of online review helpfulness and word of mouth communication on box office performance predictions. Humanit Soc Sci Commun 7, 84 (2020). https://doi.org/10.1057/s41599-020-00578-9

[12] Zhang,Y.;Bai,Z. (2023). Prediction of movies popularity in supervised learning techniques. Applied and Computational Engineering,29,142-147.

[1] The Ultimate 1Million Movies Dataset (TMDB + IMDb). 2025. Alan Vourc'h. Kaggle. https://www.kaggle.com/datasets/alanvourch/tmdb-movies-daily-updates

[2] scikit-learn developers. 2025. Understanding the decision tree structure. scikit-learn 1.8.0 documentation. [Online]. Available: https://scikit-learn.org/stable/auto_examples/tree/plot_unveil_tree_structure.html.

[3] scikit-learn developers. 2025. RandomForestClassifier. scikit-learn 1.8.0 documentation. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[4] scikit-learn developers. 2025. r2_score. scikit-learn 1.8.0 documentation. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html.

[5] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12 (Nov. 2011), 2825–2830.

[6] Python Software Foundation. 2025. statistics — Mathematical statistics functions. Python 3.14.2 documentation. [Online]. Available: https://docs.python.org/3/library/statistics.html.