International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predicting Movie Success Using Data Mining

Mr. H. Sudharsan[1], Dr. M. Iswarya[2]

[1]2nd MBA, Department of Management Science, Hindusthan College of Engineering and Technology
[2]Assistant Professor Department of Management Sciences ,Hindusthan College of Engineering and Technology, Coimbatore

Abstract:
The Tamil film industry (Kollywood) is one of the fastest-growing regional film markets in India, characterized by high investment, unpredictable audience responses, and significant financial risk. This study applies data mining and machine learning techniques to predict the box office performance of Kollywood movies, using a comprehensive dataset of films featuring actor Surya spanning from 1997 to 2025. Two machine learning models Random Forest Regression and Support Vector Regression (SVR) are trained and compared using features such as IMDb ratings, production budget, runtime, genre, director profile, and number of audience votes. The ensemble model combining both algorithms yielded strong predictive accuracy. The study further demonstrates the practical application of the model by forecasting the expected box office collection and financial outcome for the upcoming film 'Karuppu'. The findings confirm that data-driven approaches can significantly reduce financial uncertainty in film investment decisions, offering producers and investors a reliable framework for evaluating movie projects before committing capital.

Keywords: Box Office Prediction, Kollywood, Random Forest, Support Vector Regression, Data Mining, Machine Learning, IMDb, Tamil Cinema.

1. Introduction

The Tamil film industry, popularly known as Kollywood, occupies a prominent position in Indian cinema, producing over 200 films annually and attracting investment from producers, distributors, and streaming platforms. Despite the scale of this industry, decision-making around movie investment remains largely intuitive, relying on star power, director reputation, and genre trends rather than empirical analysis. This results in significant financial losses for many productions each year.

With the proliferation of digital data through platforms such as IMDb, Box Office Mojo, and social media, there is now an unprecedented opportunity to apply quantitative, data-driven approaches to evaluate the potential success of a film before its release. Machine learning algorithms, in particular, have demonstrated their effectiveness in capturing complex non-linear relationships between input features and outcome variables, making them well suited for predicting box office revenue.

This study focuses on leveraging supervised machine learning techniques specifically Random Forest Regression and Support Vector Regression (SVR) to predict box office outcomes for Kollywood films. The dataset is centered on the filmography of actor Surya, a prominent figure in Tamil cinema, whose career

spans nearly three decades and includes a wide variety of genres, budgets, and directorial collaborations. This allows for a rich, longitudinal dataset that reflects real-world variability in commercial outcomes.

The primary motivation for this research is to fill the analytical gap in the Tamil film industry where producers and investors rely on gut instinct rather than data. A reliable prediction model can assist in budget allocation, release planning, and investment risk evaluation. The study also demonstrates a practical real-world application by forecasting the box office performance of 'Karuppu', an upcoming film starring Surya, directed by RJ Balaji.

2. Statement of Problem

The film industry has a well-documented reputation for financial volatility. Despite significant financial commitments from producers, distributors, and streaming platforms, the outcome of any given film remains highly uncertain at the time of investment. The core problem is that current decision-making frameworks in Kollywood lack a systematic, data-driven basis for evaluating potential returns.

Key challenges identified in the existing literature and industry practice include:

- Decisions are primarily based on actor popularity, director track record, and industry intuition, which are subjective and inconsistent predictors of success.
- Existing datasets from platforms such as IMDb are rarely integrated with financial data and applied to predictive modeling in the Indian regional film context.
- Most predictive studies in film analytics focus on Hollywood or Bollywood, leaving a significant gap in research covering Kollywood.
- Advanced machine learning techniques such as Random Forest and SVR have not been adequately explored for Tamil cinema prediction tasks.

This study addresses these challenges by building a data driven predictive model that uses historical performance data of a leading Tamil actor to forecast both box office collection and commercial verdict for new films.

3. Objectives of the Study

1. To predict the box office collection of Kollywood movies using machine learning models by analyzing key variables including cast IMDb ratings, director profile, genre, budget, and runtime.
2. To classify movies as financial successes or failures (Blockbuster, Hit, Average, Flop) based on profitability and return on investment (ROI).
3. To develop and validate an ensemble model combining Random Forest and SVR for improved prediction accuracy.
4. To demonstrate a practical real-world application of the model by forecasting the box office performance of an upcoming Kollywood film.

4. Review of Literature

The application of data mining and machine learning to predict movie success has gained increasing scholarly attention over the past two decades. Asur and Huberman (2010) demonstrated that Twitter sentiment analysis could forecast box office revenues with notable accuracy, establishing social media as a

valid data source for film analytics. Mestyan, Yasseri, and Kertesz (2013) extended this approach by using Wikipedia page views and editing activity as predictors of movie popularity.

Li et al. (2022) employed ensemble models including Random Forest and Gradient Boosting on features such as genre, cast, and budget, using Python-based tools including scikit-learn and pandas. Zhang et al. (2024) utilized deep neural networks incorporating variables like marketing expenditure and cast popularity, achieving strong predictive performance on Hollywood datasets. Kim (2020) applied Support Vector Machines and decision trees using audience ratings data, highlighting the versatility of classical ML approaches.

Within the Indian context, Goyal, Gupta, and Agarwal (2018) compared Random Forest, XGBoost, and SVM models for Bollywood success prediction, while Upadhyay (2018) found that Multilayer Perceptron achieved the highest accuracy among a range of classifiers applied to Indian movie data. Bhadrashetty (2024) incorporated both static features such as movie credits and dynamic features such as hashtag trends to enhance prediction models.

A significant research gap exists in the application of these methodologies to Kollywood specifically. Most studies focus on globally distributed films or Bollywood productions, where data availability and audience behavior differ substantially from Tamil regional cinema. Additionally, the role of a specific actor's cumulative filmography as a longitudinal dataset for training predictive models has not been explored. This study addresses both gaps by constructing a dataset around actor Surya's career and applying advanced ML algorithms to this domain.

5. Research Methodology

5.1 Research Design

This study adopts a descriptive and analytical quantitative research design. It applies supervised machine learning models to historical film data to evaluate the predictive relationship between input features and box office outcomes. The study is empirical in nature, grounded in secondary data analysis, with model validation carried out using standard statistical and ML performance metrics.

5.2 Data Source and Sample

Secondary data was collected from the Internet Movie Database (IMDb) covering all major theatrical releases of actor Surya from 1997 to 2025, yielding a dataset of 42 films. Each film record was verified and cleaned from the raw IMDb export format. Financial data including production budgets and box office collections were supplemented from industry reports and Box Office Mojo records.

5.3 Dataset Features

Feature	Description	Type
IMDb Rating	Audience & critic rating (1–10 scale)	Numerical
Runtime (mins)	Total movie duration in minutes	Numerical
Year	Year of theatrical release	Numerical
Genres	Primary genres (Action, Drama, etc.)	Categorical
Num Votes	Number of IMDb votes received	Numerical
Directors	Director name(s)	Categorical
Budget (Cr)	Production budget in INR Crore	Numerical

Box Office (Cr)	Worldwide collection in Crore	Target Variable
Verdict	Industry verdict (Blockbuster/Hit/Flop)	Label

5.4 Data Preprocessing

The preprocessing pipeline involved five structured stages: (1) text standardization and categorical encoding of genre and director fields; (2) removal of duplicate records; (3) feature engineering including logarithmic scaling of budget and votes to reduce skewness, ratio-based features, and interaction terms such as rating-votes product; (4) outlier handling through log transformation rather than removal to preserve blockbuster data points; and (5) data type consistency validation across all columns.

5.5 Statistical Tools and Algorithms

- Random Forest Regression – an ensemble tree-based model robust to overfitting and capable of capturing non-linear feature interactions.
- Support Vector Regression (SVR) – effective for high-dimensional datasets with complex relationships between variables.
- Ensemble Model (60% RF + 40% SVR) – weighted combination for improved predictive stability.
- Evaluation Metrics: R^2 (coefficient of determination), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).
- Cross-validation: 5-fold cross-validation to assess generalizability of models.

6. Findings of the Study

6.1 Model Performance

Both machine learning models demonstrated strong predictive capabilities on the test dataset. The Random Forest model achieved a higher R^2 score, indicating that it explained a greater proportion of the variance in box office collections. The SVR model, while slightly lower in R^2, exhibited lower Mean Absolute Error, suggesting more consistent predictions across the range of film budgets. The ensemble model, combining both algorithms with a 60:40 weighting, yielded the most balanced results across all performance metrics.

Metric	Random Forest	SVR	Ensemble
R^2 (Test)	0.82	0.78	0.84
MAE (Cr)	28.40	24.10	22.90
RMSE (Cr)	45.20	48.60	43.80
MAPE (%)	19.3	21.7	17.8
CV R^2 Mean	0.79	0.74	0.81

6.2 Feature Importance

Feature importance analysis from the Random Forest model revealed that IMDb Rating was the most influential predictor of box office collection, followed by Production Budget, Number of Votes, and Director's historical box office average. Genre also contributed significantly, with Action and Drama genres demonstrating the strongest positive correlation with commercial performance. Runtime showed a moderate

positive relationship, suggesting that audiences in Kollywood tend to respond favorably to full-length feature films above 140 minutes.

6.3 Career Trajectory Analysis

Analysis of Surya's box office timeline from 1997 to 2025 reveals three distinct performance eras: an initial phase (1997–2004) characterized by moderate collections, a peak commercial phase (2005–2015) driven by blockbusters such as 'Kaakha Kaakha' and 'Singam', and a recent diversification phase (2016–2025) featuring varied genres and budget scales. The data shows a strong positive correlation between high IMDb ratings and box office success, validating audience quality perception as a financial driver.

6.4 Karuppu Box Office Prediction

The trained ensemble model was applied to forecast the commercial performance of 'Karuppu', an upcoming Surya film directed by RJ Balaji with an estimated production budget of ₹130 Crore. Input parameters included an assumed IMDb rating of 7.2, an estimated 18,000 audience votes, and an Action-Drama genre classification.

Model	dicted BO (Cr)	Net Profit (Cr)	ROI (%)	Expected Verdict
Random Forest	₹214.50	₹84.50	65.0%	Hit ✓
SVR	₹198.30	₹68.30	52.5%	Hit ✓
Ensemble (60/40)	₹208.30	₹78.30	60.2%	Hit ✓

The ensemble model predicts that 'Karuppu' is expected to achieve a box office collection of approximately ₹208 Crore against a budget of ₹130 Crore, yielding an estimated ROI of 60.2% and placing the film in the 'Hit' category. This outcome is consistent across all three model variants, lending confidence to the prediction.

7. Implications and Recommendations

For Producers and Investors

The model provides a pre-production risk assessment tool that can guide investment decisions based on quantifiable input parameters. Producers should prioritize projects with director profiles that have consistent box office track records, as director historical average emerged as a significant predictor. High-budget projects should be evaluated through the model to ensure sufficient revenue probability before committing to full production expenditure.

For the Film Industry

Production houses should invest in structured data collection and digital record-keeping for all productions, including budget breakdowns, marketing expenditures, and audience engagement metrics. Standardized reporting, similar to the BRSR framework in the financial sector, would significantly enhance the quality and consistency of future predictive models.

For Streaming Platforms

Streaming platforms such as Amazon Prime, Netflix, and Sun NXT can leverage similar predictive models to determine optimal acquisition prices for Kollywood films based on expected box office performance.

OTT rights valuations could be tied to model-predicted revenue tiers, reducing negotiation uncertainty.

For Policy and Academic Research

The study demonstrates the viability of applying standard ML methodologies to Indian regional cinema. Future research should expand the dataset to cover multiple actors, multiple regional film industries, and integrate social media sentiment data as real-time dynamic features.

8. Conclusion

This study demonstrates that machine learning-based predictive models can serve as effective decision-support tools for the Kollywood film industry. By applying Random Forest Regression and Support Vector Regression to a dataset of 42 films from actor Surya's filmography (1997–2025), the study establishes a quantitative framework for forecasting box office collections and financial outcomes.

The key findings confirm that IMDb rating, production budget, director track record, and genre are the most significant predictors of commercial success in Tamil cinema. The ensemble model, combining Random Forest and SVR at a 60:40 ratio, achieved the strongest overall performance with an R^2 of 0.84 and a MAPE of 17.8%. The practical application of the model to forecast the upcoming film 'Karuppu' yielded a consistent prediction of a 'Hit' verdict across all model variants.

These findings highlight that data-driven decision-making is both feasible and valuable in the context of Indian regional cinema. Reducing reliance on intuition and increasing the use of analytical tools can materially lower financial risk for producers and investors, supporting a more sustainable and strategically managed film industry. Future work should incorporate real-time social media sentiment, trailer engagement metrics, and multi-actor datasets to further improve predictive accuracy and generalizability.

References

1. Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
2. Mestyan, M., Yasseri, T., & Kertesz, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLOS ONE, 8(8).
3. Li, X., et al. (2022). Predicting Movie Success with Machine Learning Approaches. Applied Sciences, 12(4), 1981.
4. Zhang, Y., et al. (2024). Movie Success Prediction Using Neural Networks. Journal of Data Mining and Digital Humanities.
5. Kim, J. (2020). Machine Learning Methods for Movie Popularity Prediction. International Journal of Information Technology, 12(3), 745–751.
6. Goyal, A., Gupta, P., & Agarwal, S. (2018). Bollywood Movie Success Prediction Using Machine Learning. International Journal of Engineering and Technology, 7(3), 302–307.
7. Upadhyay, A. (2018). Movie Success Prediction Using Data Mining. International Journal of Computer Applications, 181(12), 1–5.
8. Bhadrashetty, A. (2024). Movie Success and Rating Prediction Using Data Mining Techniques. Journal of Computer Science and Technology.
9. Kudagamage, C., et al. (2019). Data Mining for Movie Success Analysis and Prediction.

Proceedings of the International Conference on Data Science and Engineering.
10. Darapaneni, N. (2020). Movie Success Prediction Using Machine Learning. WEKA-based Analysis. International Journal of Advanced Research in Engineering and Technology.
11. Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts. Management Science, 53(6), 881–893.
12. Mishra, A., et al. (2021). Sentiment Analysis in Movie Revenue Prediction. International Journal of Natural Language Processing, 8(2), 33–41.