IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Movie ROI Prediction Using Explainable Machine Learning And Decision Support System

[1]Vedula Jahnavi, [2]Anurag Shivam

[1]Student, [2]Student

[1]School of Computing,

[1]SASTRA Deemed University, Thanjavur, India

Abstract: Movie production includes huge amount of financial risk, and investment decisions are frequently based on previous experience or market intuition rather than statistics. We propose a machine learning based methodology for predicting a movie's Return on Investment (ROI) based in information available before its release. Our proposed model considers budget, runtime, genre, director, production company, release timing and movie overview to estimate the expected ROI. Multiple machine learning models were compared, and Random Forest showed the best performance and it was selected as final prediction model. Then, we used SHAP (Shapley Additive exPlanations) to find and explain the factors that significantly impact each prediction. A Decision Support System (DSS) is developed to help producers estimate different production scenarios before making investment decisions. This allows the user to change important factors such as budget, runtime, franchise status and release timing and see how those changes affect the predicted ROI. Our system suggests an investment based on the estimated ROI and helps producers in making better investment decisions.

Index Terms – Movie ROI Prediction, Machine Learning, Random Forest, SHAP, Explainable Artificial Intelligence (XAI), Decision Support System (DSS), Return on Investment (ROI), Film Industry.

I. INTRODUCTION

Film production requires a huge amount of financial investment but still there is no guarantee of success. A movie's performance is affected by many factors including audience preferences, marketing, competing releases, reviews and time of release. This leads to the producers depending on their intuition when deciding about their investment. This method is not reliable enough to compare different production alternatives and make investment decisions.

Previously, machine learning algorithms have been used to predict movie ratings, box-office revenue, or predict whether a movie will be a hit or a flop. Such projections are useful, but they do not give producers the kind of information they require the most. Producers are more interested in knowing the expected Return on Investment (ROI), the level of investment risk, and the reasons behind the prediction. Existing studies rarely provide a way to test different production scenarios before making an investment decision.

For solving these problems, our research presents an explainable machine learning model for ROI prediction of movies using only the pre-release information. While existing models mainly focus on the prediction problem, our proposed framework combines prediction, explainability based on SHAP, and the Decision Support System (DSS) that enables producers analyze various production options before investing.

The main contributions of this paper are as follows:

- Developed a machine learning system to predict movie ROI using information available before the movie is released.
- Tested different regression models to find the best one for predicting movie ROI.
- Used SHAP to understand why the model gives certain predictions.
- Created a Decision Support System to help producers test different movie plans and decide whether investing in a movie is a good idea.

II. Related Work

Many studies have used machine learning to predict different factors of movie performance. Past studies, however, mainly focused on ratings, revenue, or movie success. Very few studies predict Return on Investment (ROI), explain the predictions or help producers in deciding where to invest. [1] developed a machine learning model for predicting movie ratings and revenue using historical data and Random Forest. However, the study does not support investment decisions. [2] predicted OTT content performance using ratings and audience interaction. The results showed that higher ratings were related to better content performance. However, the study is limited to success prediction but not explanation. [3] compared different machine learning models to predict movie revenue using pre-release information. It did not predict profitability or explain the model's predictions. [4] focused on movie revenue using information available before release, however it didn't explain the prediction made by the model. [5] proposed a model to predict a movie's success before release, but focused only on success and not on financial outcomes. [6] used data mining techniques on historical data to classify movies as hit, average or flop. Although the approach allowed early prediction of movie success, it did not explain its predictions. [7] combined social media data with SHAP to predict movie success before release and identify the factors influencing the predictions. This mainly focused on performance prediction and did not provide any investment recommendations. [8] used Random Forest Regression with ensemble learning to predict movie ratings using TMDB dataset. This improved rating prediction using better feature encoding. [9] used machine learning and SHAP to identify how production budget, release season and popularity affect a movie's revenue. The study identified the important factors affecting revenue, but did not provide producers a framework to improve their investment decisions. [10] developed a machine learning model called Cine-Predict that predicts a movie's success before release using metadata. The model only predicts the movie's success and does not give recommendations to producers for improving it. [11] used Random Forest to predict IMDb ratings using pre-release information from various sources. While this supports production planning, it is limited to predicting ratings and does not provide decision support. [12] used machine learning techniques to predict movie profit using metadata and review information, but does not explain their predictions or provide recommendations for producers. The literature survey shows that existing studies mainly focus on movie-related predictions such as ratings, revenue, profitability, or overall success using machine learning. As many recent studies have used explainable AI, few have combined ROI prediction, model explainability, and a decision Support System into a single framework. This work proposes an explainable machine learning framework that predicts movie ROI, explains the predictions using SHAP, and integrates a Decision Support System that allows what-if analysis for better investment choices before production.

III. RESEARCH METHODOLOGY

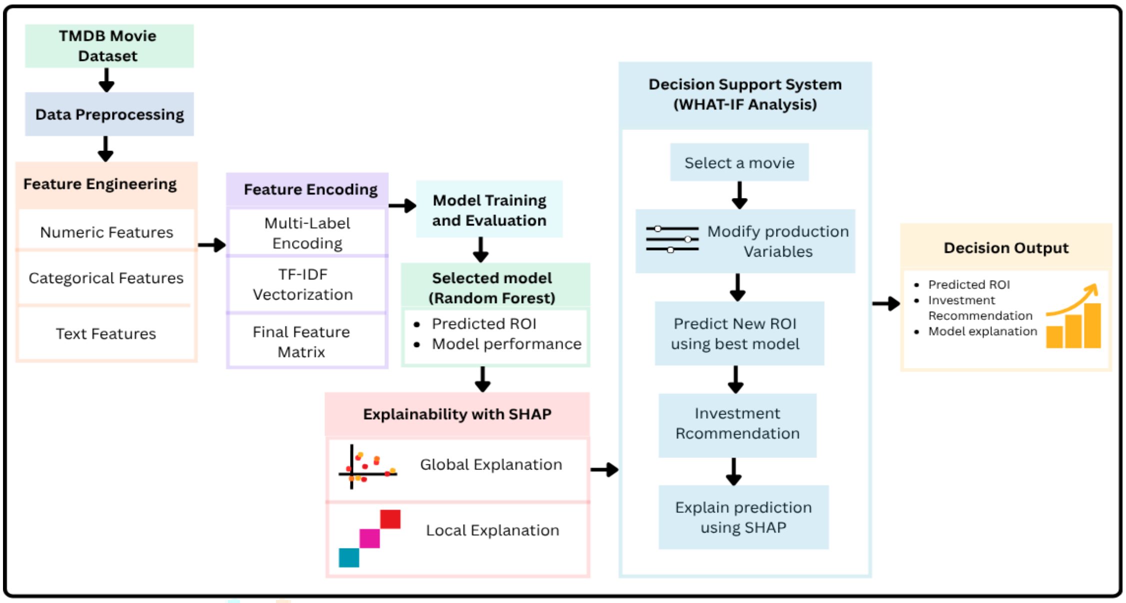

Figure 1: workflow of the proposed movie ROI prediction framework and Decision Support System

3.1 Population and Sample

The Movies Dataset was used in this study. The movie details, cast and crew records, and keyword information are combined into a single row for each movie using a common identifier across the source files. A consolidated table was made containing each movie's financial information, production details, genre, crew and description, which forms the input workflow illustrated in Fig. 1.

3.2 Data Preprocessing

The combined dataset was cleaned and ROI was calculated which was used as the target variable. The missing values were handled based on the nature of the data. The missing values related to the franchise were considered as non-franchise movies, while records with missing or invalid budget and revenue values were dropped as ROI could not be computed for those rows. The budget and revenue were converted to numeric form and from these values profit and ROI were calculated using Eq. 1 and Eq. 2. Extreme outliers were removed to minimise their impact on training and evaluation of the model.

$$\text{Profit} = \text{Revenue} - \text{Budget} \quad (1)$$

$$\text{ROI} = (\text{Revenue} - \text{Budget}) / \text{Budget} \quad (2)$$

3.3 Feature Engineering

Additional features believed to influence ROI were derived. Timing features such as release year, month, quarter, season and franchise indicator were derived. Categorical information such as genre, director, production company, language and country was extracted from the raw data, and directors with very few films were grouped into a single "Other" category to avoid the model from overfitting to categories with almost no data. The movie overview was retained as a text feature. These features were then converted into numerical values for model training.

3.4 Feature Encoding

Every categorical, multi-label and text feature was converted into numeric form. Genres, keywords, production companies, countries and spoken languages were multi-label encoded, while original_language, release_season and the cleaned director field were one-hot encoded. The synopsis was transformed into numerical vectors using TF-IDF after removing English stop words. All encoded blocks were merged with the numeric features to produce a final feature matrix, which was then split into training and test sets.

3.5 Machine Learning Model Development

Multiple regression models were trained and compared to identify the most suitable approach for predicting movie ROI. The dataset was split into a training set (80%) and a testing set (20%). Since the ROI of a movie is affected by many financial, production and textual factors that may be linearly and non-linearly related, models from different learning approaches were evaluated. The linear baseline was Ridge Regression, the tree-based methods were Decision Tree and Random Forest, and the gradient-boosting methods were

XGBoost, LightGBM and CatBoost. All models were trained with the same feature set evaluated by MAE, RMSE, R^2 and training time. Fig. 2 Illustrates the procedure of model development and selection. Based on the comparative evaluation results, the Random Forest model was selected for the explainability analysis and development of the Decision Support System.

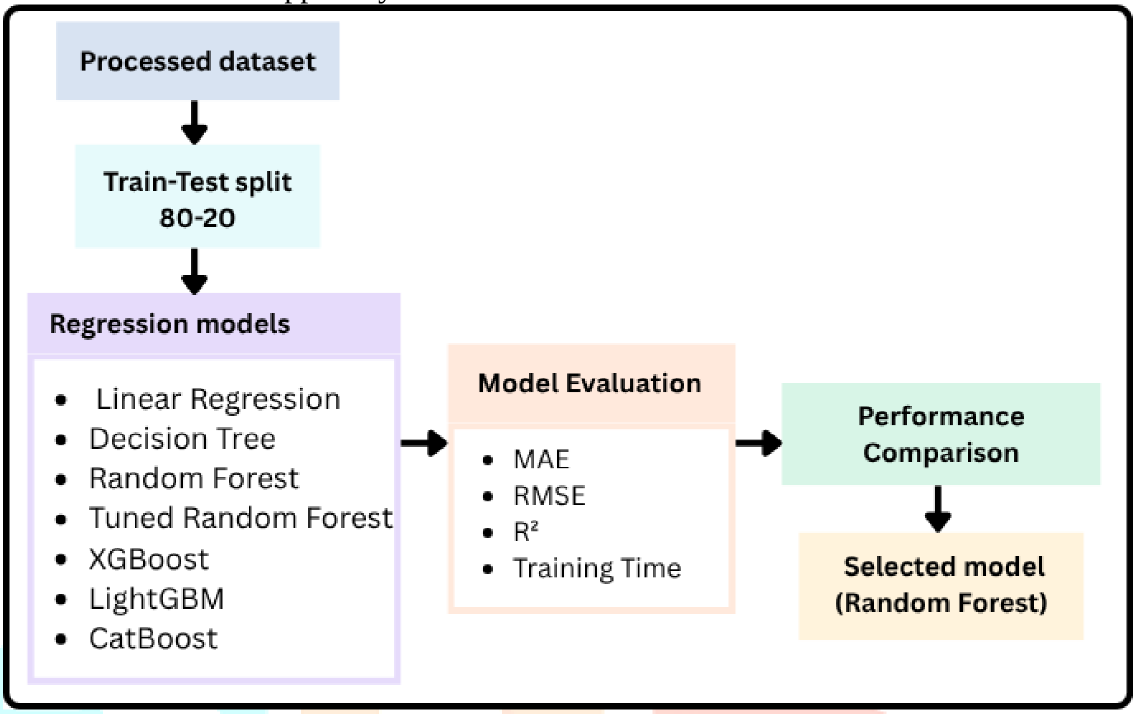

Figure 2: Machine learning model training, evaluation, and selection process.

3.6 Explainable AI

The trained model's predictions were interpreted at both the dataset level and the individual-movie level. A SHAP explainer was applied to the trained Random Forest model, to improve the transparency of the predicted ROI. A SHAP summary plot was generated to provide a global, dataset-wide view of feature impact, while waterfall plots were generated for three representative movies (high, average and low ROI) for a local, case-by-case view. The overall SHAP workflow is illustrated in Fig. 3.

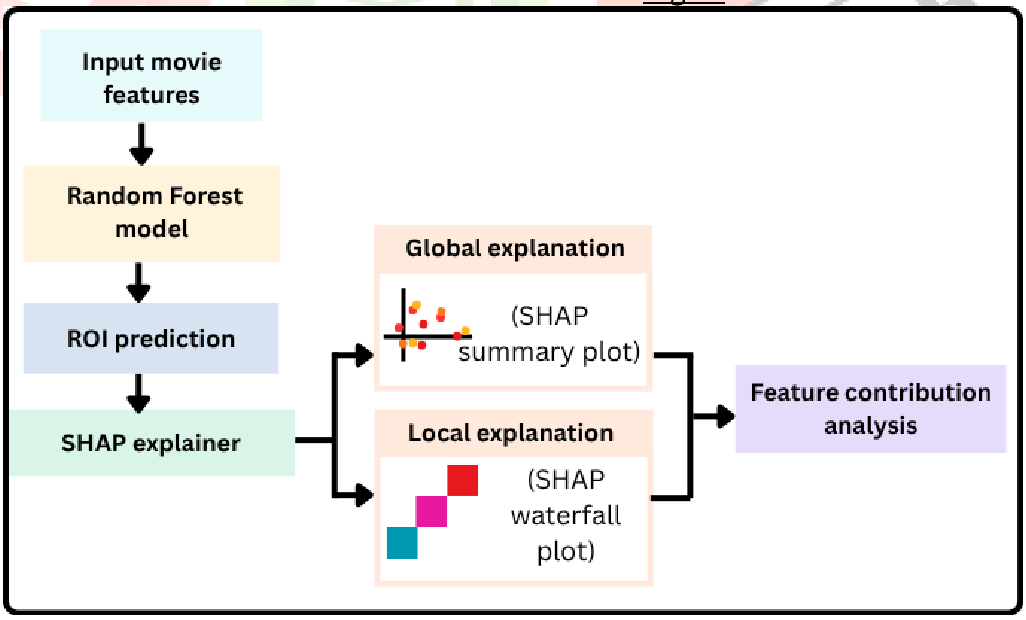

Figure 3: Workflow of SHAP-based global and local explanation for movie ROI prediction.

3.7 Decision Support System

The Decision Support System (DSS) developed in this study turns the trained model into a what-if analysis tool for producers, since a single static ROI prediction is only useful if it can inform a decision. The trained Random Forest model, along with the pre-processed feature components, is loaded into the DSS, where a producer selects a base movie and modifies parameters such as budget, runtime, franchise status, release year and release month, while all other engineered features are kept fixed at their original values. The modified inputs are passed through the model to generate a new predicted ROI, which is compared with the movie's original ROI and classified into a risk level and investment recommendation (ROI < 0 → "Do Not Invest", High Risk; 0 ≤ ROI < 1 → "Invest with Caution", Medium Risk; ROI ≥ 1 → "Recommended Investment", Low Risk), and re-explained using SHAP as shown in Fig. 4.

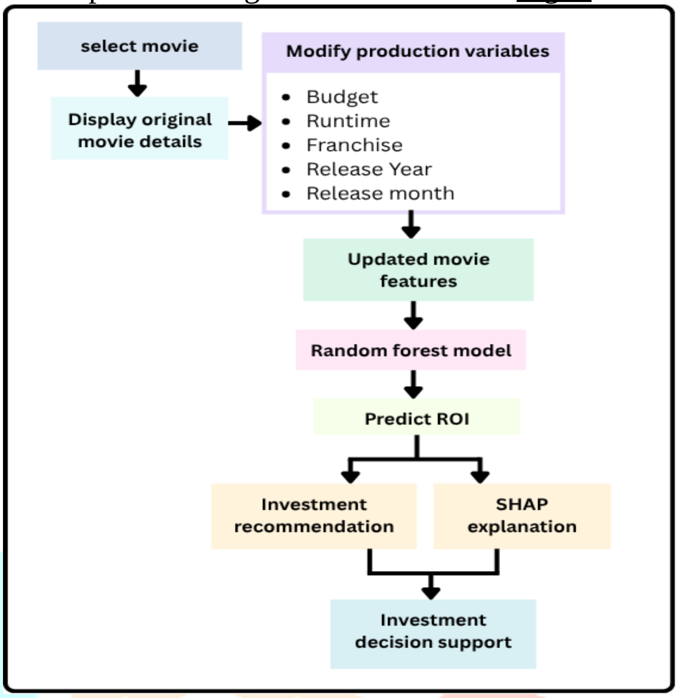

Figure 4: Workflow of the Decision Support System for what-if analysis and investment recommendation.

IV. RESULTS AND DISCUSSION

4.1 Model Performance

Table 1: Performance comparison of machine learning models for movie ROI prediction.

Model	MAE	RMSE	R^2	Training Time (s)
Random Forest	3.0862	6.1811	0.1872	14.83
LightGBM	3.2550	6.2689	0.1639	1.82
XGBoost	3.2252	6.2795	0.1611	7.05
CatBoost	3.2222	6.3498	0.1422	17.73
Tuned Random Forest	3.2826	6.4708	0.1092	46.25
Decision Tree	3.3176	6.6669	0.0544	0.65
Ridge Regression	4.7927	7.2905	-0.1308	0.77

Table 1 shows a comparison of the performance of all seven regression models. Random Forest has achieved highest value of R^2 (0.1872), lowest MAE (3.0862), and RMSE (6.1811). LightGBM and XGBoost gave similar results, but Ridge Regression had the weakest performance since its R^2 value was negative, showing that movie ROI does not follow a simple linear relationship. The moderate R^2 achieved by Random Forest implies that the pre-release data can only explain some of the variance in ROI, while important post-release factors such as marketing, critic reviews, audience reception, and competing releases were not available to the model. Therefore, Random Forest was selected for the further SHAP analysis and DSS development.

4.2 Feature Importance

The feature importance scores for the Random Forest model are shown in Fig. 5. Budget was the most important feature, followed by release year, franchise, runtime and production company. These variables contributed the largest share of the model's overall importance, whereas genre, director, and overview-based text features had comparatively smaller individual contributions. A few TF-IDF keywords were ranked in the top features, but their impact was far below that of the main production-related variables. Overall, the financial and production variables play greater role in predicting the ROI than any other textual variable.

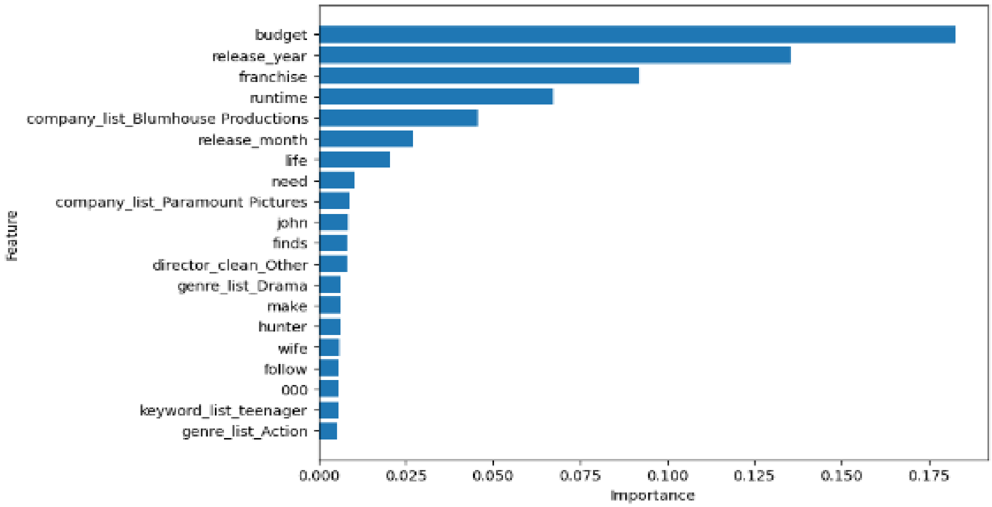

Figure 5: Top 20 most important features identified by the Random Forest model.

4.3 SHAP Local Explanation

SHAP explanations were generated for three representative movies (high, average and low ROI). The prediction increases from a baseline value of 2.92 to a final predicted ROI of 3.86. Franchise status provides the largest positive contribution (+2.22) whereas budget (-0.59) and release year (-0.54) reduce the predicted ROI. The remaining features make comparatively smaller positive and negative contributions to the final prediction. The waterfall plot provides a transparent explanation of how each feature influences the predicted ROI for an individual movie.

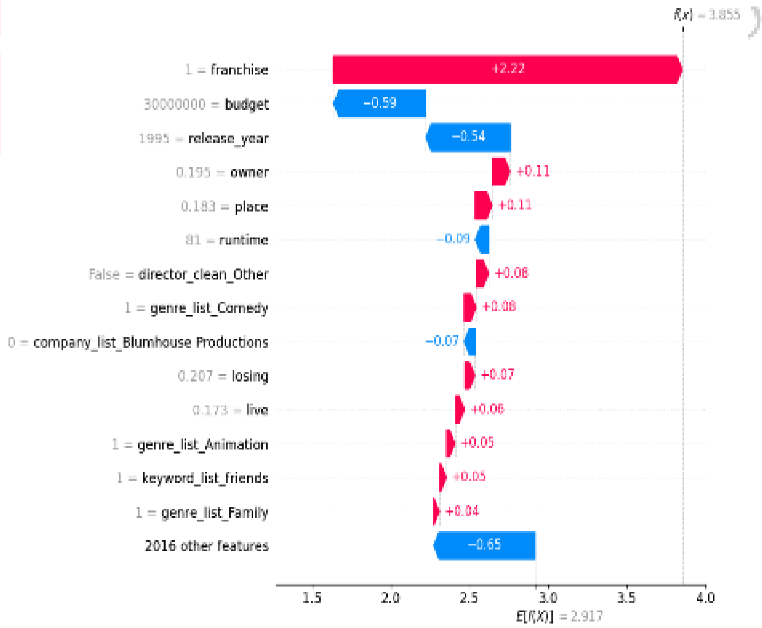

Figure 6: Local explanation of an individual movie prediction using SHAP.

4.4 SHAP Global Explanation

The SHAP summary plot for the global impact of features on movie ROI prediction, where each point represents one movie, while the colour indicates whether the feature value is high or low. Franchise status, budget, release year and runtime are always the most significant for ROI prediction. Higher franchise values are generally associated with positive SHAP values, whereas higher budget values frequently contribute negatively to predicted ROI. The summary plot indicates that the model is dependent on some key production factors and does not depend on single instances from the dataset.

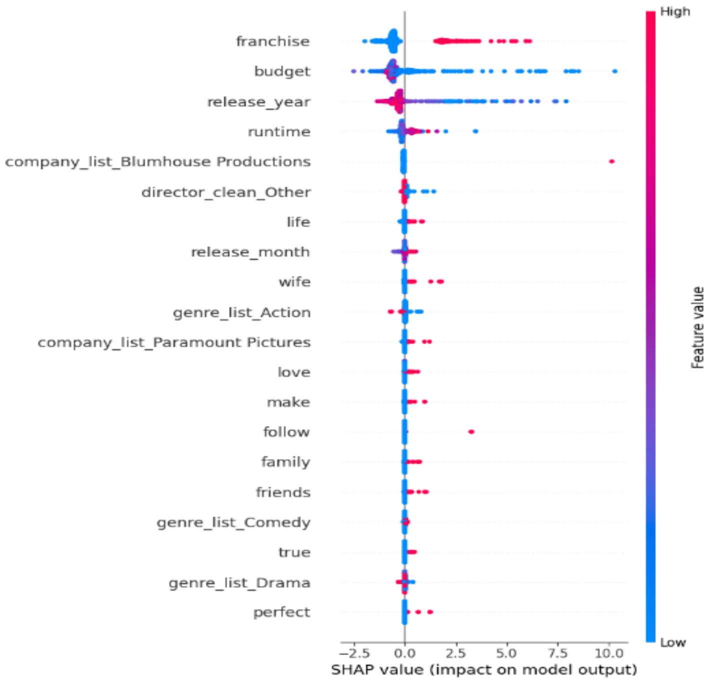

Figure 7: Global feature importance and impact on movie ROI prediction using SHAP.

V. Conclusion

This proposed methodology is an explainable machine learning framework that predicts the Return on Investment (ROI) of a movie before its release. This framework uses information about the movie that is available during the planning stage to estimate ROI. It also explains the prediction using SHAP. It supports investment decisions through a Decision Support System. Many machine learning models were tested, and Random Forest had the best overall performance. Although predicting movie ROI is a difficult task because many important factors become available only after release, the proposed model was able to capture meaningful patterns from the information available before release. SHAP helped in identifying the features that contributed most to each prediction. The Decision Support System demonstrates the practical value of the proposed framework by allowing producers to modify important production parameters and observe how these changes affect the predicted ROI and investment recommendation. This study shows that explainable machine learning can be a decision-support tool for producers before they invest their money. This work can be extended in future by using larger datasets, as there are other factors such as actor popularity, marketing budget, audience reviews and social media trends for improving the prediction performance of the model. A web-based application can be created for the Decision Support System, it will allow producers to evaluate various production scenarios in a simple way and plans real-time investments.

REFERENCES

[1] Singh, K.K., Makhania, J. and Mahapatra, M., 2024. Impact of ratings of content on OTT platforms and prediction of its success rate. *Multimedia Tools and Applications*, *83*(2), pp.4791-4808.
[2] Bomma, R.S., Singhal, S., Bose, S.C., Rahul, G. and Koyyalamudi, J., 2025, October. The Role of Ratings in OTT Content Success: A Machine Learning Approach. In *2025 3rd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)* (Vol. 1, pp. 876-882). IEEE.
[3] Udandarao, V. and Gupta, P., 2024. Movie revenue prediction using machine learning models. *arXiv preprint arXiv:2405.11651*.
[4] Paul, C. and Das, P.K., 2022. Predicting movie revenue before committing significant investments. *Journal of Media Economics*, *34*(2), pp.63-90.
[5] Hoque, M.A. and Khan, M.M., 2022, September. Machine Learning Approaches to Predict Movie Success. In *International Conference on Machine Intelligence and Emerging Technologies* (pp. 613-627). Cham: Springer Nature Switzerland.
[6] Gopinath, P., Saini, P., Kumar, N. and Meel, P., 2025, October. Movie success prediction using data mining. In *AIP Conference Proceedings* (Vol. 3297, No. 1, p. 020001). AIP Publishing LLC.
[7] Abdulrashid, I., Ahmad, I.S., Musa, A. and Khalafalla, M., 2025. Impact of social media posts' characteristics on movie performance prior to release: an explainable machine learning approach. *Electronic Commerce Research*, *25*(6), pp.4533-4557.
[8] Marpid, N.N., Kurniawan, Y.I. and Rahayu, S.P., 2025. Analysis Of the Movie Database Film Rating Prediction with Ensemble Learning Using Random Forest Regression Method. *Jurnal Teknik Informatika (JUTIF)*, *6*(1), pp.1-10.
[9] Torkamani, M.J., Gomes, P., Sadeghnejad, A. and Le, J., 2026. Analyzing the Impact of Release Season and Production Budget on Movie Revenue and Profitability. *arXiv preprint arXiv:2605.12551*.
[10] Lyngdoh, J., 2025, June. Cine-Predict: A Movie Success Predicting system. In *2025 IEEE Guwahati Subsection Conference (GCON)* (pp. 1-5). IEEE.
[11] Hathboor, S.A., 2025. *Predicting Hollywood Movie Success Using Predictive Machine Learning Algorithms*. Rochester Institute of Technology.
[12] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20). Cini, K., 2025. *Analysing script features for performance forecasting in film production: a data-driven approach leveraging NLP* (Master's thesis, University of Malta).