**Box Office Prediction Using Machine Learning Techniques**

**A report on**

**Box Office Prediction Using Machine Learning Techniques**

**[CSE- 3183]**

Submitted By

**Boya Yashasvi Naidu 210962070**

**V Abhiram Varma 210962072**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**MANIPAL INSTITUTE OF TECHNOLOGY,**

**MANIPAL ACADEMY OF HIGHER EDUCATION**

**NOV 2023**

***Abstract— In the dynamic and multi-billion-dollar film industry, accurately predicting a movie's box office revenue is crucial for making informed investment decisions. The proposed machine learning model integrates various parameters, such as production company, genre, budget, reviews, and ratings, to forecast a movie's revenue based on pre-release information. This predictive tool aims to provide investors with intelligent insights, enabling them to navigate the complexities of the film business and minimize risks. Additionally, the study addresses the evolving field of predicting societal responses to new products, specifically within the motion picture industry. The introduced decision support system, leveraging machine learning techniques and historical data from platforms like IMDb, Rotten Tomatoes, Box Office Mojo, and Metacritic, assists investors in approximating a movie's success rate and profitability, contributing valuable insights to mitigate risks in the ever-changing landscape of the film industry.***

***Keywords*— Box Office, Linear Regression; KNeigbors, SVM; Random Forest; Machine Learning**

1. INTRODUCTION

The film industry's potential for substantial revenue, exemplified by India's 2019 box office revenue of Rs 10,948 crore, attracts significant investments. The film industry is a vast domain of investment opportunities, but also marked by complexity and significant risks associated with predicting box office success. With the exponential growth of available movie data, analysing this field becomes compelling. However, predicting a movie's success is intricate, influenced by diverse revenue streams and differing definitions of success. Technological advancements provide a wealth of data, yet predicting box office success remains complex. Success is often measured by revenue, crucial for investors and stakeholders. Therefore, predictive models become invaluable, aiding investment decisions and informing production strategies for potential success.

This study focuses on predicting a film's revenue from a production point of view. Machine learning algorithms like Linear Regression, KNeighbors Regressor, Decision Trees, Random Forest, SVM. The paper delineates previous research, data descriptions, methodology, and experimental outcomes, offering insights into predicting box office success.

The concept underpinning this project revolves around creating models to predict a movie's pre-release box office revenue. Such models would benefit investors, aiding decision-making, and support production companies in strategizing before commencing film production.

1. LITERATURE REVIEW

[1] The paper predicts movie box office success using SVM and Neural Network on a dataset of 755 movies from 2012-2015, considering pre-release features like reviews, ratings, and star power. Using a five-class classification system, it assesses accuracy, favoring neural networks. The conclusion suggests incorporating audience-related factors and genre for improved predictions. [2] The project focused on Creating a Box Office Success Recommendation System with machine learning, using TMDB and OMDB APIs. Features extracted, outliers removed, and regression models (Ridge, Random Forest, XGBoost, CatBoost, LGBM) analyzed. The Voting Regressor stood out, predicting revenue based on genre, uncovering insights into budget, runtime, star power, and expected popularity. A reverse regression model accurately estimated features for desired revenue. Promising results, future tweaks could involve more feature interaction and exploration of advanced models like CNNs, RNNs, or GANs. [3] Explores regression algorithms for predicting movie revenue using IMDb and Facebook metadata. Aims to optimize ad spending by forecasting success. Merges datasets, handles revenue disparities, and engineers features. Backward elimination for feature selection based on p-values. Models compared: Linear, SVR, Decision Tree, Random Forest, Ridge, Lasso. Random Forest outperforms, identifying key predictors as vote\_count and budget. Results hint at ML success in revenue prediction, with Random Forest leading. Future work may involve integrating social media data for real-time improvements. [4] Aimed to predict movie success via IMDb data using a web scraper. Preprocessing involved deduplication, handling missing values, and filtering low-rated or low-revenue movies. Features: title length, release date, genre, actor/director/writer ratings. Reduced dataset via sequential feature selection. Used decision trees, SVM, KNN for revenue and rating predictions. Moderate success in ratings (65-83%), struggled with revenue (15%). Challenges: data limitations, duplicates, genre preprocessing. Future work may include expanding data sources and refining preprocessing. [5] Predicts box office using linear models and classification. Linear models: naive and modified regression, locally weighted regression, focusing on opening weekend box office. Classification: SVM with Naive Bayes, neural network categorizing films into 10 groups by gross box office. Results favor modified and locally weighted regression. Classification shows promise, needs improvement. Future: add plot-related features, explore SVM-linear regression combos. [6] Proposes a model to enhance movie success prediction using various KNN algorithms. Methodology involves data gathering from IMDB, cleansing, feature selection, and modeling. Fine, Weighted, Medium, and Cubic KNNs outperform Cosine and Crude KNNs. Emphasizes optimizing KNN algorithms for accurate predictions. Future work includes refining the model with additional data sources. [7] Analyzes social media and IMDB data to predict movie success, using SVM and Neural Networks. Finds neural networks outperform SVM, emphasizing the impact of sequel success, actor popularity, and social media sentiment on movie outcomes. Recommends expanding features and platforms for better prediction accuracy in future work. [8] Predicts Bollywood movie success using machine learning, emphasizing star cast, genre, director, and budget. Utilizes web-scraped data, employing models like linear regression, logistic regression, naive bayes, SVM, and k-means, with linear regression demonstrating the highest accuracy. Suggests improving accuracy by including social media parameters and expanding predictions to web series and other media.

III. RESEARCH GAP AND OBJECTIVES

Integration of Diverse Data Sources: Existing box office prediction models predominantly focus on factors like genre, cast, and marketing budget. However, integrating additional sources such as audience sentiment analysis from social media, historical box office performance, present political scenario, critical reviews, etc can result in more comprehensive models. These integrated sources would offer a more holistic understanding of a movie's potential success.

Granular Prediction Models: Current models primarily forecast overall box office revenue, but there's a gap in predicting specific performance aspects like opening weekend revenue, international box office collections, or DVD/online sales. Developing granular models that predict these specific revenue streams could provide nuanced insights valuable for film studios and distributors.

Attribute Analysis for Box Office Success: This study aims to analyse correlations between various movie attributes (e.g., release date, director, franchise association) and box office performance metrics. Positive or negative correlations can unveil the significant impact of specific attributes on a movie's box office success, guiding filmmakers, and studios in decision-making processes.

Accurate Prediction Models: The primary goal is to create predictive models or methodologies that accurately forecast box office performance based on intrinsic movie attributes. This involves employing machine learning algorithms or statistical techniques to develop models closely aligned with actual box office earnings, aiding in reliable revenue predictions.

Identification of Key Predictive Attributes: Understanding which specific movie attributes significantly influence box office success is crucial. The objectives include pinpointing key factors (such as budget, cast, popularity etc.) strongly correlated with box office revenue, providing actionable insights for filmmakers and production houses.

IV. METHODOLOGY

The algorithms used in this paper were Logistic Regression, KNeighbors Regressor, Decision Trees Regressor, SVR and Random Forest Regressor after splitting the dataset in the ratio of 70:30.

A diagram of a data processing process

Description automatically generated

Fig 1: Block Diagram of the entire methodology

*A. Data Acquisition:*

The dataset utilized in this study comprises of details of movies from all around the world while the majority of the movies are from the U.S. The dataset is acquired from the global TMDB dataset.

For the experiments conducted in this paper, a subset. consisting of 3000 movies was extracted, from the larger dataset. The relevant features included in this study are as follows:

“id”, “belongs\_to\_collection”, “budget”, “genres”, “homepage”, “imdb\_id”, “original\_language”, “original\_title”, “overview”, “popularity”, “poster\_path”, “production\_companies”, “production\_countries”, “release\_date”, “runtime”, “spoken\_languages”, “status”, “tagline”, “title”, “Keywords”, “cast”, “crew”, “revenue”.

These attributes serve as the foundation for predicting the revenue of a movie.

*B. Preprocessing and Feature Extraction:*

The dataset underwent modifications involving essential feature updates. Missing values labeled as 'na' in numerical columns such as Budget and runtime were substituted with the mean value of their respective columns. For categorical columns like Genre, Production countries, and spoken languages, missing 'na' values were replaced with the mode of those specific columns. Additionally, a collection was created for each item not associated with any collection. Counts for Genre, Production companies, production countries, and spoken languages were calculated. Following the removal of all NULL values from entries, the 'release date' feature was modified by being split into three distinct features, for the day, month, and year of the release.

*C. Data Exploration:*

A graph with blue dots

Description automatically generated

Fig 2: Genre counts vs revenue

As seen from the plot above, movies with 3 genres in general make the highest revenue.

A graph with blue dots

Description automatically generated

Fig 3: Languages counts vs revenue

From the plot, it is observed that single language films are more likely to make more revenue.

A line graph with different colored lines

Description automatically generated

Fig 4: Release month vs mean of revenue, budget, popularity

Observing the trend between the release month and revenue, budget, popularity it can be noted that every category peaks in the 6th month. This is understandable as that is the summer season and many well-known blockbusters are summer releases. People as well as production companies would prefer watching and releasing movies in the summer respectively.

*D. Data Splitting:*

Divide the pre-processed dataset into training and testing sets using a 70:30 ratio. The training set (70%) will be used for model training, while the testing set (30%) will evaluate model performance.

*E. Model Development:*

The model development process involved implementing and training five distinct machine learning algorithms: Linear Regression, KNeighbors Regressor, Decision Trees, Random Forest Regressor, and SVR, with the aim of identifying the most effective method for predicting box office revenue.

Linear Regression Implementation:

Linear Regression, a foundational statistical technique, was employed to establish relationships between various factors and box office revenue. This model aimed to predict revenue based on independent variables such as budgets, release dates, cast etc. While assuming a linear relationship between predictors and revenue, this method provided insights into the potential impact of these factors on box office success.

KNeighbors Regressor Implementation:

The KNeighbors Regressor, employing proximity-based learning, predicted box office revenue by identifying similarities between movies based on features such as genre, cast, and production budget. By considering the revenue of the k-nearest movies, this model leveraged similar film characteristics for accurate revenue estimations.

Decision Trees Implementation:

Decision Trees, a non-linear model, were implemented to predict box office revenues by creating a hierarchical structure based on various movie attributes like genre, director, and release month etc. By iteratively partitioning the data, this model aimed to capture nonlinear relationships, offering insights into which features significantly impact box office success.

Random Forest Regressor Implementation:

Utilizing an ensemble approach, Random Forest Regressor was applied to predict box office revenues by aggregating multiple decision trees. By combining diverse predictions from individual trees, this model aimed to enhance accuracy and generalizability, considering various movie characteristics, marketing strategies, and historical data to provide more robust revenue forecasts.

SVR (Support Vector Regression) Implementation:

Support Vector Regression, capable of handling both linear and non-linear relationships, was implemented to predict box office revenue by finding an optimal hyperplane in a higher-dimensional space. By considering movie attributes and their impact on revenue, while optimizing parameters like C (regularization), kernel type, this model aimed to accurately predict box office earnings by identifying intricate patterns within the data.

IV. DISCUSSION AND ANALYSIS OF RESULTS

From the correlation matrix obtained, it can be seen that, none of the attributes are highly correlated to each other (excludi. So, no attribute will be removed.

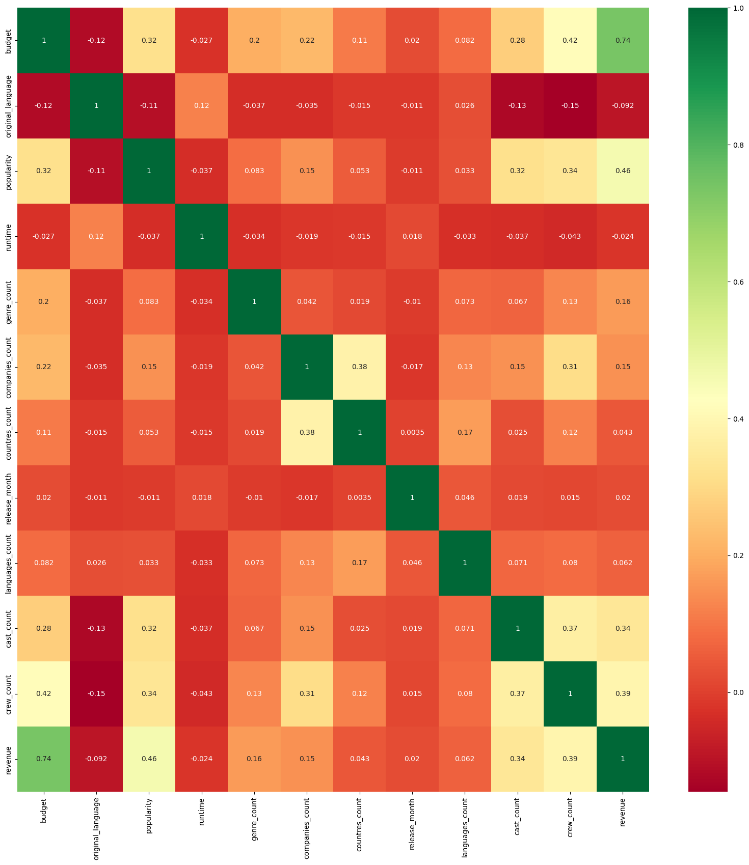


Fig 5: Correlation heatmap of attributes

The decision to employ Linear Regression, KNeighbors Regressor, Decision Trees, Random Forest Regressor, and Support Vector Regression (SVR) for predicting box office revenue comes from their varied characteristics, making them suitable for predictive modeling in the film industry. Linear Regression offers a fundamental linear approach, KNeighbors Regressor captures local patterns, Decision Trees enable rule-based predictions, Random Forest Regressor utilizes ensemble learning for robustness, and SVR handles non-linear relationships, providing a diverse toolkit for accurate box office revenue forecasts.

*A. Parameter Tuning:*

For optimizing predictive models, a meticulous selection process ensues, involving linear regression, support vector machines, k-nearest neighbor, random forest, and Decision Trees. Each of these models harbors a set of hyperparameters, pivotal in sculpting their performance landscapes. The crux lies not only in implementing these models but also in properly navigating their hyperparameter spaces to reveal their optimal configurations.

Tuning these parameters guides towards successful model outputs. Through this method, the goal is not just to apply these algorithms but to also fine-tune their settings, unlocking their full potential and extracting the most accurate, insightful predictions from each model.

For Linear Regression, employing a grid search can enhance predictive performance by tuning parameters such as normalization, intercept, and regularization strength.

Tuning the KNeighbors Regressor includes exploring the optimal number of neighbors, the distance metric, and weights to balance bias and variance effectively.

In Decision Trees, focusing on parameters like maximum depth, minimum samples for splitting, and decision-making criteria aids in hyperparameter optimization.

The Random Forest Regressor benefits from fine-tuning the number of trees, maximum depth, and feature selection criteria to balance diversity and predictive accuracy. It is observed from varying the n\_estimators from 100 – 1000, the accuracy only improved by 0.66%.

For Support Vector Regression (SVR), hyperparameter optimization involves tuning kernel selection, regularization parameter (C), and kernel coefficient (gamma) to capture intricate data relationships, contributing to more accurate predictions.

IV. CONCLUSION AND FUTURE WORK

|  |  |
| --- | --- |
| Model | Accuracy |
| Linear Regression | 62.14 |
| KNeighbors Regressor | 59.81 |
| Decision Tree Regressor | 35.62 |
| Random Forest Regressor | 68.73 |
| SVR | -15.48 |

Table 1: Accuracies of each model

From the table, it can be seen that the random forest model achieves the highest accuracy at 68.73%. This suggests its adeptness in comprehending complex relationships between features and the target variable. Compared to other models, it demonstrates superior generalization to unseen data.

Although the remaining models display decent accuracy, the linear regression model comes closest to the random forest's performance, achieving 62.14%. Nevertheless, their accuracy doesn't match the random forest's level, indicating its superiority for this dataset and task.

The model could be further developed to be able to predict the post-release revenue using factors like public talk, number of days ran, etc. It would be interesting to develop a hybrid model that predicts the revenue based on all the input factors and could also give the necessary features required to attain a said input revenue.

REFERENCES

[1] N. Quader, M. O. Gani, D. Chaki and M. H. Ali, "A machine learning approach to predict movie box-office success," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2017, pp. 1-7, doi: 10.1109/ICCITECHN.2017.8281839.

[2] G. Velingkar, R. Varadarajan, S. Lanka and A. K. M, "Movie Box-Office Success Prediction Using Machine Learning," 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2022, pp. 1-6, doi: 10.1109/ICPC2T53885.2022.9776798.

[3] Faaez Razeen “Predicting Movie Success using Regression Techniques

[4] Pojan Shahrivar and Carl Jernbacker “Predicting Movie Success using Machine Learning Techniques” Stockholm, Sverige 2017

[5] Pengda Liu “Machine Learning on predicting gross box office” 2016

[6] Ajmed Al Fahoum and Tahani A. Ghobon “Accurate Machine Learning Predictions of Sci-Fi Film Performance” Journal of New Media 2023

[7] Midhun Josey “Predicting Box Office Success of Movies Using Social Media” 2019

[8] Ansari Sana Fatima and Shruti Pimple “Movie Success Prediction Using Machine Learning” 2020 pp. 10-14