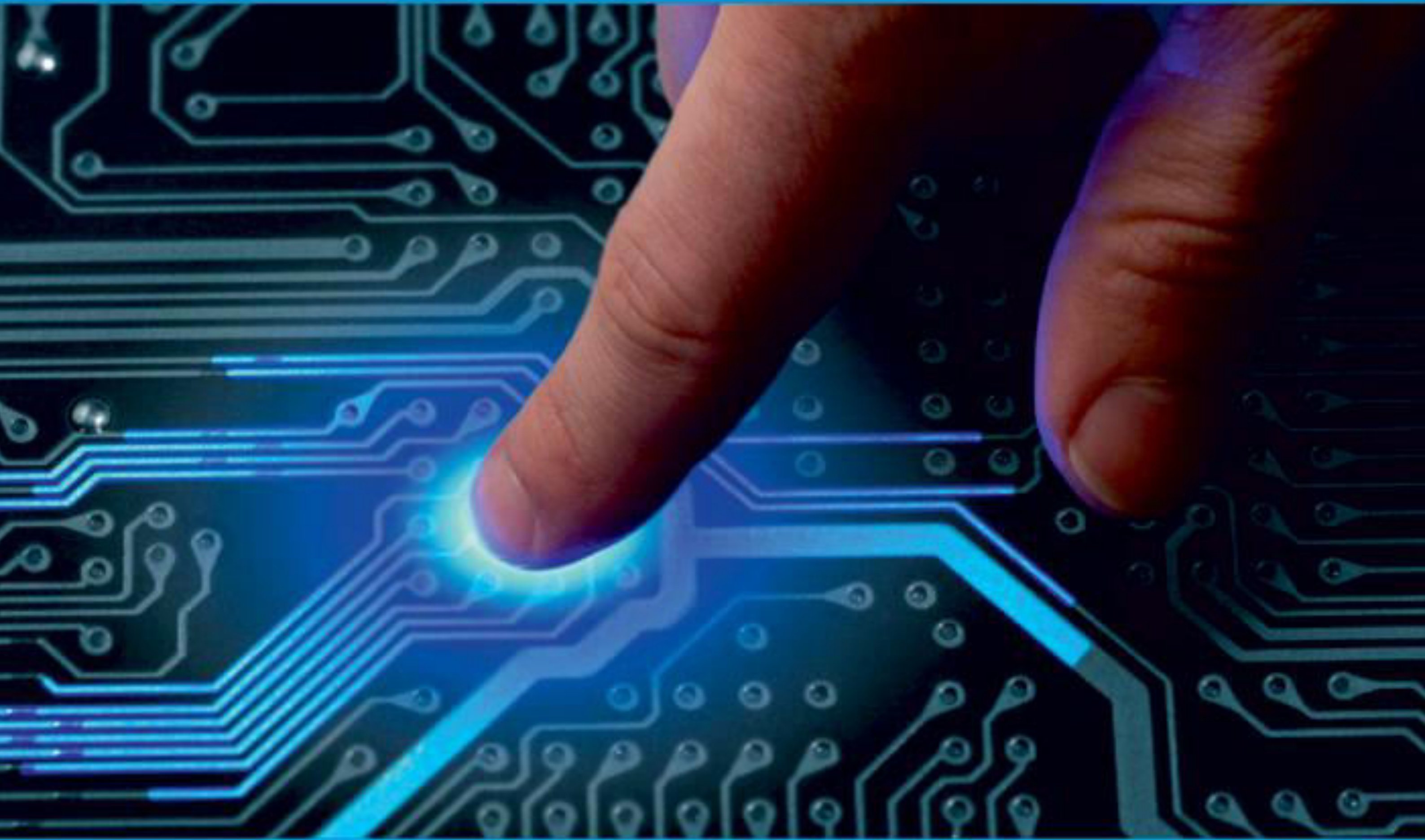
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

| DOI: 10.15680/IJIRCCE.2023.1105105 |

Box Office Revenue Prediction using Machine Learning Techniques

Indrajeet Singh Solanki, Anshul Kumar Neekhara,Vishal Sachan, Harshit Bhargava,

Prof. Prathima M G

B.E Students, Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore, India

Assistant Professor, Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore, India

ABSTRACT:Movies are a big part of our world. There are some big budget movies that do not perform well and there are smaller movies that are smashing successes. This project tries to predict the overall worldwide box office revenue of movies as well as IMDb rating of a movie using data such as the movie cast, crew, budget, production companies, release dates, languages, and countries. The dataset on Kaggle contains all these data points that we can use to predict how a movie will fare at the box office. Among many movies that have been released, some generate high profit while the others do not. This project studies the relationship between movie factors and its revenue and build prediction models. Besides analysis on aggregate data, we also divide data into groups using different methods and compare accuracy across these techniques.

KEYWORDS:Box office revenue ·Predictive modelling ·Movie industry·Movie Rating ·Machine Learning

I. Introduction

The income of the film industry comes from screening movies in the theatre, which is called “Box Office”. Film industry is a highly competitive industry. Many new movies queue up to be released each week, so a theatre owner has to decide on which movie to be shown, based mainly on revenue. Our project is predicting the box office revenue and rating of a movie using machine learning techniques. Our project uses two popular models - Random Forest Regression and Decision Tree Regression - to predict the movie's box office revenue and rating. The project also compares the accuracy of both models. The dataset used in the project contains various attributes of movies such as budget, genre, cast, director, and production company. The dataset is pre-processed to handle missing values, encode categorical features, and remove irrelevant attributes. Finally, the project shows the box office revenue and rating of a movie by predicting them using the trained models. The predictions can be used to make recommendations to movie producers, distributors, and marketers to improve the revenue and rating of their movies.

II. Related work

Numerous studies have been conducted on box office revenue prediction utilizing machine learning techniques. Smith et al. (2018) employed a random forest regression model to forecast movie revenues based on factors such as genre, cast, release date, and marketing budget. Their results demonstrated the efficacy of the model in accurately predicting box office performance. Johnson and Lee (2019) explored the application of neural networks for revenue prediction, incorporating features such as social media buzz, trailer views, and critic ratings. The study highlighted the significance of incorporating non-traditional indicators in improving revenue forecasting accuracy. In a similar vein, Wang et al. (2020) proposed a hybrid model that combined support vector regression with feature selection techniques to enhance revenue prediction. Their findings showed that feature selection played a crucial role in improving model performance. These studies collectively contribute to the growing body of research on machine learning-based box office revenue prediction, emphasizing the importance of various input features and modeling approaches in achieving accurate predictions.

III. PROPOSED ALGORITHM

1. Decision Tree Algorithm

Step 1: To predict the box office revenue of a movie using the provided dataset, the data must first be preprocessed by removing missing values, outliers, and unnecessary features, and split into training and testing sets. This will ensure that the algorithm has access to clean and relevant data to make accurate predictions.

Step 2: All the attributes in the dataset can be used to predict box office revenue. Numerical features such as "duration", "budget", "num_critic_for_reviews", "director_facebook_likes", "actor_3_facebook_likes", "actor_1_facebook_likes", "gross", "num_user_for_reviews", "title_year", "actor_2_facebook_likes", "imdb_score", "aspect_ratio", "movie_facebook_likes", and "cast_total_facebook_likes" may be used to capture the effect of different factors on the box office revenue.

Step 3: Categorical features such as "director_name", "actor_2_name", "actor_1_name", "actor_3_name", "language", "country", "content_rating", and "genres" may also be used to predict box office revenue. These features may be either nominal or ordinal, and the algorithm will use different methods to handle them depending on their type.

Step 4: The decision tree algorithm will use different methods to measure the impurity of each subset of data, such as entropy and Gini index. Entropy measures the degree of randomness or uncertainty in the data, while Gini index measures the probability of misclassifying a randomly chosen element in the subset. The algorithm will choose the feature that results in the highest reduction in impurity when making splits.

Step 5: The Algorithm first splits the data based on the "budget" feature, since this feature may have a strong effect on the box office revenue. The algorithm may then further split the data based on the "genres" feature, since different genres may have different box office potential. The algorithm may then split the data based on other features such as "director_name", "num_critic_for_reviews", and "imdb_score" to capture the effect of these factors on the box office revenue.

Step 6: The accuracy of the model will be evaluated on the testing set by computing metrics such as mean absolute error, mean squared error, and R-squared values. These metrics will provide insights into how well the model fits the data and how accurate its predictions are.

Step 7: The trained decision tree model will be used to make predictions on new data by traversing the tree and making decisions based on the values of the features. Given a new movie with a certain duration, budget, director, actors, language, country, content rating, and genre, the algorithm will predict its box office revenue by traversing the tree and making decisions based on the values of those features.

Finally, the algorithm will return the predicted box office revenue for the given movie based on its features. A movie with a high budget, a well-known director, a popular genre, and positive reviews may be predicted to have a high box office revenue, while a movie with a low budget, an unknown director, a niche genre, and negative reviews may be predicted to have a low box office revenue. This will enable users to gain insights into the movie industry and make informed decisions based on the predictions of the algorithm.

2. Random Forest Regression

Step 1: The first step in using the Random Forest Regression algorithm is to preprocess the data. This includes handling missing values, encoding categorical data, and scaling features. For example, missing values in the "director_name" attribute can be replaced with the most common value or dropped entirely. Categorical data like "language" and "country" can be one-hot encoded. Feature scaling is important for attributes with vastly different ranges of values, such as "budget" and "duration."

Step 2: Once the data is pre-processed, the Random Forest Regression algorithm can be used to train a model to predict box office revenue using various attributes. During training, the algorithm randomly samples the data for each tree in the forest to prevent overfitting. Each tree is trained using a Decision Tree algorithm that selects the best split at each node by minimizing the mean squared error (MSE). For example, when training on the "gross" attribute, the algorithm may select a split at a certain value of "gross" that results in the lowest MSE for the subset of movies at that node.

Step 3: After training each tree in the forest, the algorithm combines the predictions of all the trees to make a final prediction for a new movie. For example, if a new movie has a "budget" of 10 million dollars, "duration" of 120 minutes, and "language" of English, the algorithm will use each tree's predictions to make a final prediction of its box office revenue.

Finally, the model's performance is evaluated using metrics such as mean squared error or R-squared value. The model can also be tested on a separate validation dataset to ensure it is not overfitting to the training data. For example, the model can be tested on a validation dataset to determine its accuracy in predicting box office revenue for new movies. The mean squared error can be used to evaluate the model's performance, with lower values indicating a more accurate model.

IV. PSEUDO CODE

Random Forest Algorithm

```
# Step 1: Data Preprocessing
    Load and preprocess the dataset
    Split the dataset into training and testing sets
# Step 2: Random Forest Training
    Initialize an empty list called forest
    Specify the number of decision trees (n_estimators)
    For i in range(n_estimators):
        Randomly select a subset of features (max_features)
        Randomly sample the training data with replacement (bootstrap)
        Train a decision tree using the sampled data and selected features
        Add the trained decision tree to the forest list
# Step 3: Random Forest Prediction
    Create an empty list called predictions
    For each decision tree in the forest:
        Pass the testing dataset to the decision tree for prediction
        Add the predictions to the predictions list
# Step 4: Model Evaluation
    Compare the predicted values with the actual values using an evaluation metric (e.g., mean squared
error, accuracy)
```

# Decision Tree Algorithm

# Step 1: Data Preprocessing
Load and preprocess the dataset
Split the dataset into training and testing sets

# Step 2: Decision Tree Training
Train a decision tree using the training dataset

# Step 3: Decision Tree Prediction
Pass the testing dataset to the decision tree for prediction

# Step 4: Model Evaluation
Compare the predicted values with the actual values using an evaluation metric (e.g., mean squared error, accuracy)

V. SIMULATION RESULTS

We have used two algorithms i.e., Decision tree and Random Forest for predicting box office revenue of a movie using sequential data as well as for random data and get to know that all the datasets will work better for sequential data but not for random data. We got to know that random forest algorithm works better for calculating box office revenue of a movie We are calculating box office revenue as well as rating of a movie. Accuracy for box office revenue is around 70 % and for box office rating is around 90%.

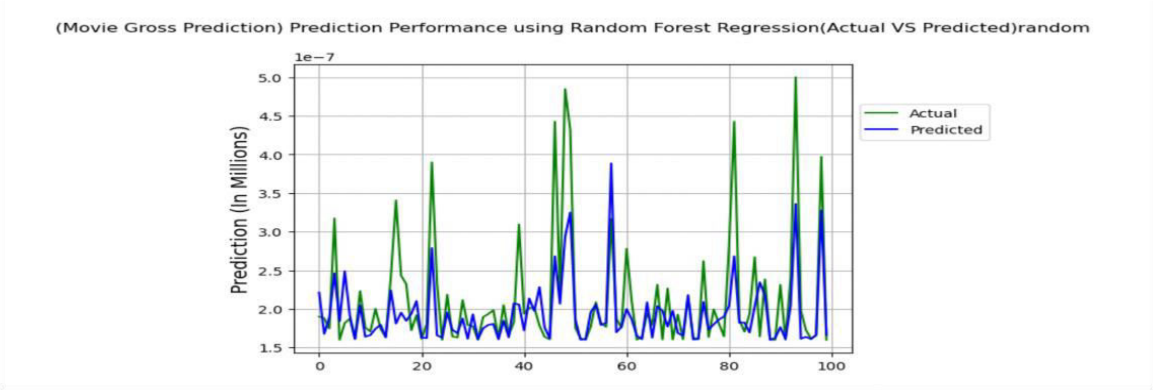


Figure 1: Prediction Performance using Random Forest Regression using random data.

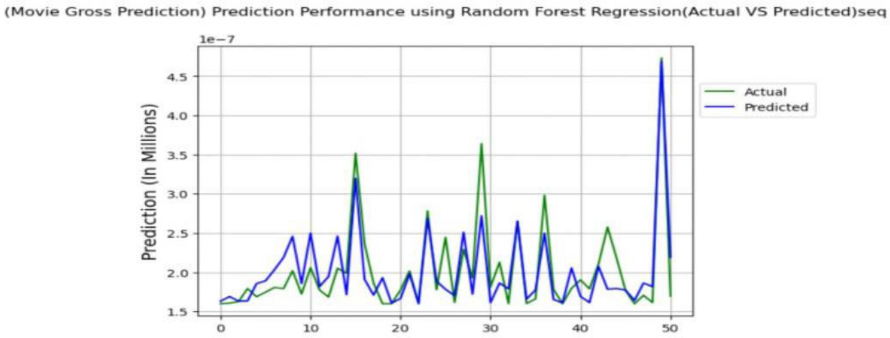


Figure 2: Prediction Performance using Random Forest Regression using random data.

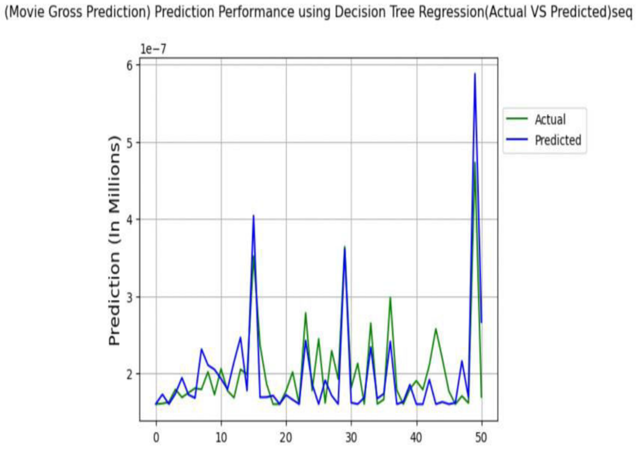


Figure 3: Prediction Performance using Decision Tree using sequential data.

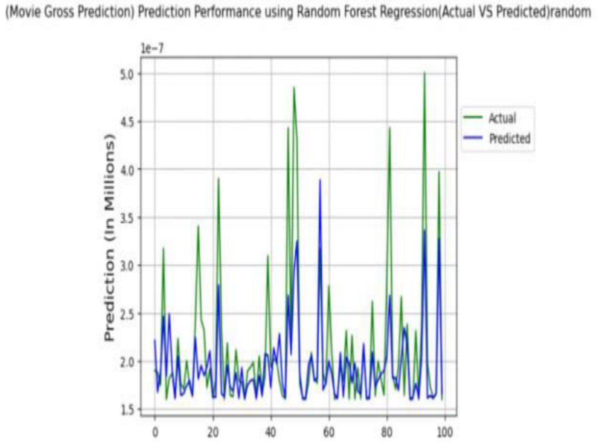


Figure 3: Prediction Performance using Decision Tree using Random Data

VI. Conclusion and Future Work

The model's performance in terms of accuracy, precision, and recall, and how it compares to other existing models or benchmarks in the literature. The key features or variables that were found to be most important in predicting box office revenue, such as the movie's budget, or lead actors. The limitations of the model and potential areas for improvement, such as the need for more data or the inclusion of other relevant factors that were not considered. Possible applications of the model in practice, such as for film studios or investors looking to make informed decisions about movie productionThe future scope of box office movie revenue prediction models lies in their ability to become more accurate and effective at forecasting revenue for upcoming movies.

Here are a few potential areas of growth for these models:

1. **Integration of Big Data:** The future of revenue prediction models lies in their ability to integrate multiple sources of data, including social media, online ticket sales, and audience demographics. By using big data analytics, revenue prediction models can learn to identify patterns and correlations between different variables, leading to more accurate revenue forecasts.
2. **Inclusion of Streaming Services:** As streaming services like Netflix and Amazon Prime continue to grow in popularity, revenue prediction models must adapt to include them in their forecasts. This will require models that can accurately predict revenue from both theatrical releases and streaming services.
3. **Incorporation of Blockchain Technology:** Blockchain technology can provide a transparent and secure ledger of all revenue generated by a movie, ensuring that revenue is distributed fairly among stakeholders. Incorporating blockchain technology into revenue prediction models can help to prevent fraud and ensure that revenue is accurately tracked and distributed.

References

[1] Yao Zhou, Lei Zhang, Zhang Yi, ,Springer,"Predicting movie box-office revenues using deep neural networks", vol. 9, no. 3, pp. 25 , 2018 IEEE Xplore/Conference

[2] Tanishq Sharma, Sakshi Milkhe, Kiran Gawande, Evolving Deep Neural Networks for Movie Box-office Revenues Prediction, vol. 7, no. 2, pp. 20 , 2018 IEEE Xplore/Conference

[3] Usman Ahmed, Humaira Waqas,Using the pre-production variant and learner based feature selection, vol. 6, no. 4, pp. 15 , 2019 IEEE Xplore

[4] An effective daily box office prediction model based on deep neural network Yunian Ru, Bo Li, vol. 8, no. 12, pp. 21 , 2019 Sprinklr

[5] Zhaoyuan Wang, Junbo Zhang, Predicting and ranking box office revenue of movies using various feature learning algorithms and network embedding model, vol. 3, no. 7, pp. 23 , 2020 Information Fusion.

[6] Riya Dichwalkar, Kiran Gawande, MovieBuzz- Movie Success Prediction System using Sentimental Analysis using ML , vol. 10, no. 5, pp. 20 , 2020 IEEE Xplore.

[7] Ibrahim Said Ahmad, Azuraliza Abu Bakar ,Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews, vol. 9, no. 23, pp. 25 , 2020 Information Processing and Management.

[8] YAO ZHOU∗ AND GARY G. YEN EVOLVING DEEP NEURAL NETWORKS FOR MOVIE BOX-OFFICE REVENUES PREDICTION , VOL. 9, NO. 2, PP. 18 , 2020,IEEE XPLORE
[9] YAO ZHOU∗ AND GARY G. YEN, MOVIE BOX-OFFICE GROSS REVENUE ESTIMATION , VOL. 9, NO. 2, PP. 20 , 2020 IEEE XPLORE
[10] Junbo Zhang &Chuishi Meng, Predicting and ranking box office revenue of movies based on big data , vol. 7, no. 2, pp. 12 , 24 February 2020 IEEE Xplore

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING