Long–Range Movie Box Office Prediction Based on Machine Learning

Jiayi Xu *

Department of Physical Science, University of California, Irvine, USA

* Corresponding author: jxu16@uci.edu

Abstract. The study aims to predict the box office of movies based on several important factors obtained from The Movie Data Base (TMDB) being independent of the daily box office of the movie after its release. Exploring the nature of the dynamics of film revenue is crucial for filmmakers, investors, and film likers since it can help them find a way to improve their film influences and decision-making. This research employs three distinct methods – Random Forest, Back Propagation Neural Network, and Linear regression (Least Square Method) – utilizing a set of selected independent variables including popularity, budget, genre, run time, release date, original language, and production countries. This study gives an insight into the relationships between these variables and the revenue of movies. As a long-range forecasting model, the prediction of the model is up to 73% precise. Overall, this study aims to provide valuable information and methods for the film-related industry to predict revenue. Also, the model has the potential to be extended to other fields such as the prediction of the feedback of series or games before release.

Keywords: Box office prediction, machine learning, random forest, neural network.

1. Introduction

Have you ever wondered what factors contribute to the blockbuster success of a film or the financial challenges faced by filmmakers in an ever-evolving entertainment industry? The ability to accurately predict the revenue of a film not only intrigues film enthusiasts but holds immense practical value for filmmakers, studios, and investors. However, the existing long-range forecast's accuracy is far below that of the short-range forecast.

In today's fast-paced and competitive film industry, where production costs are soaring, and audience preferences are dynamic, the ability to forecast the financial performance of a film before it is released is more than a mere academic exercise. It is a strategic necessity that can influence critical decisions in production, marketing, and distribution.

The film industry is characterized by high levels of financial risk, with production costs often reaching astronomical figures. Filmmakers and studios grapple with the challenge of predicting audience reception, market trends, and the myriad factors that contribute to a film's commercial success. As such, the value of a robust predictive model for movie revenue cannot be overstated.

Against this backdrop, this study delves into the intricate task of predicting the revenue of a film, exploring the relationships between various key variables. By employing advanced predictive models and analyzing factors such as budget, popularity, release timing, and genre.

2. Literature Review

2.1. Research Objective

Nowadays, the most accurate movie box office models predict the revenue after the movie is released based on the data such as feedback from a test screening or the box office of the opening week, with less uncertainty and reasonably more accuracy [1]. The primary objective of the research is to develop a model to predict the total revenue of movies before release, which is a long – ranged forecast. Using a compilation of machine learning predictive approaches like random forest, BP neural network, and least squares linear regression, the research aims to achieve the targeted objective. A collection of significant parameters, including budget, release date, runtime, genre, production

country, original language, and popularity, will be used for training and evaluating our chosen models. Aside from that, the results of models will mainly be evaluated by R square and thus those methods can be compared and evaluated under this topic.

2.2. Method Introduction

This research utilizes three models – Random Forest, LSE Linear regression, and BP neural network to predict the revenue of the movie.

2.2.1. Random Forest

Random Forest Regression is the method of regression involving creating decision trees.

The formation of each tree refers to randomly selecting instances and feature variables from the training data set. Each tree produces criteria and evaluative measures closely corresponding to the samples.

Finally, the Random Forest Algorithm finishes its regression by utilizing all the criteria and estimations from each decision tree within the forest [2].

2.2.2. Least Square Estimation Linear Regression

Linear regression is a method in mathematical statistics to employs regression analysis to find how two or more variables, which have one independent variable included, are influencing each other linearly in terms of quantity.

As the method assumed, the final relationship between those variables will be approximated by a straight line [3].

2.2.3. Back Propagation Neural Network

The BP neural network is a type of layered forward network trained using the backward error propagation algorithm, and it is currently one of the most popular models among applied neural networks.

The BP neural network mainly learns by dynamically toning the weights and thresholds of the networks through the process of backpropagation to minimize the sum of squared errors [4].

3. Methodology

3.1. Data Acquisition

The data is extracted from The Movie Data Base (TMDB), including 3000 rows and 23 columns. From those 23 variables, some, such as IMDB ID, the link to the poster, and the overview of the movie, are either useless or hard to quantify. Thus, I only extracted 7 kinds of value from them: revenue (the dependent variable), budget, popularity, runtime, release date, original language, and production countries.

3.2. Data Exploratory Analysis

3.2.1. Standardisation

Figure 1 and figure 2 are the histogram of the revenue and budget, it is obvious that the data is significantly skewed.

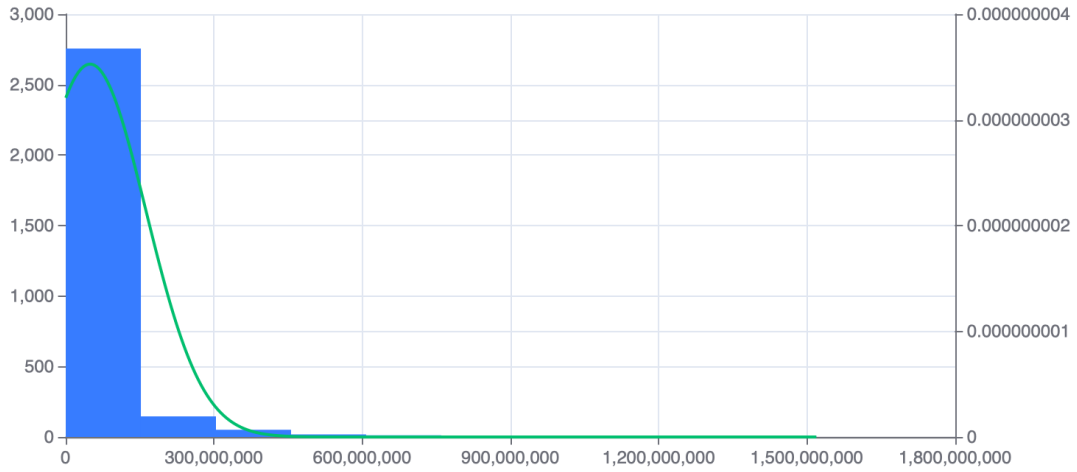

Figure 1. Quantity versus Revenue

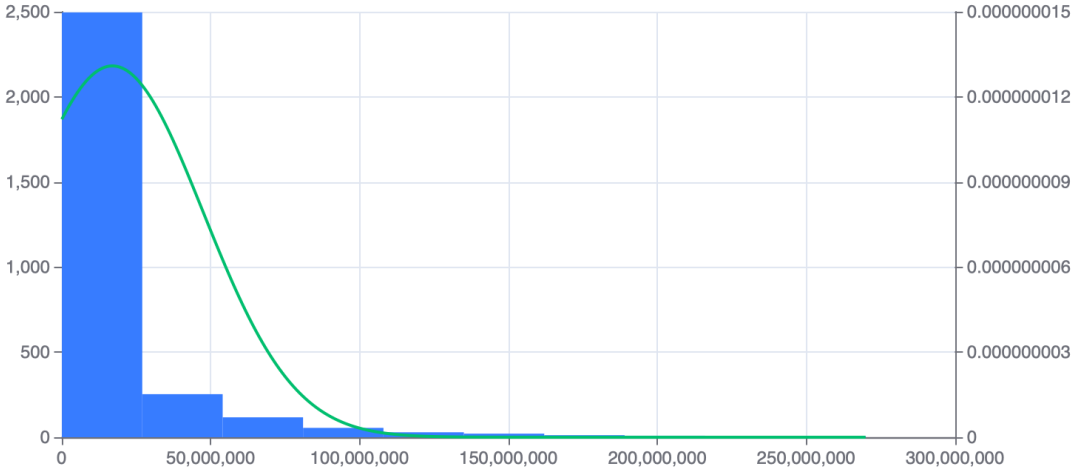

Figure 2. Quantity versus Budget

Revenue, budget, popularity, and runtime, the numerical features, are all skewed and not distributed normally. In addition, the quantity levels of different features are distinct. Thus, to decrease the negative influence of processing the data in LSE linear regression and BP neural network, which are the models sensitive to unnormalized and distinctive data, this paper should standardize the data.

Standardization of the data by the min-max standardization method can ensure that the data in different units and magnitude can be unified which enhances the performance of the modeling method of Linear Regression and BP Neural Network.

Min-max standardization:

$$X' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

3.2.2. Release Date

This research breaks down the release date into two parts – release years and release months -- to explore the seasonal changes of the movie and how its performance evolves over those years.

3.2.3. Original Language and Production Countries

The original language and production countries imply how regional and cultural factors influence the movie box office.

3.2.4. Genre

This study extracts the primary genre of the movie, and this study want to see how the category of the movie influences the revenue.

3.2.5. Correlation

This paper uses Kendall's tau-b method to determine the correlation between different variables.

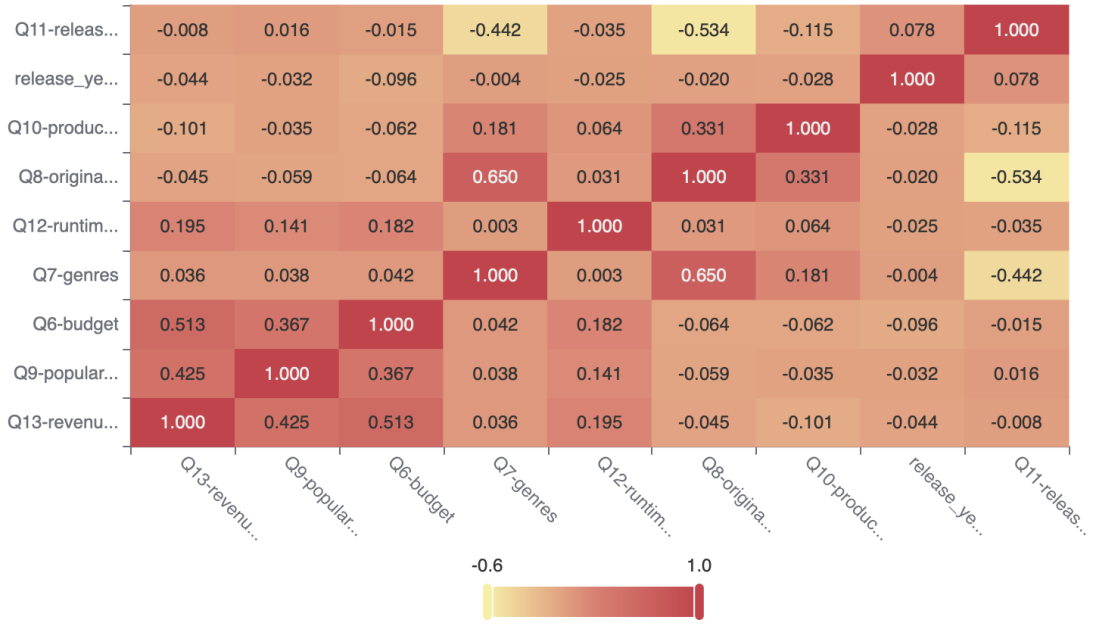

Figure 3. Correlation Heat Map

As shown in figure 3, popularity, budget, and runtime exhibit strong correlations with movie revenue. Movies with higher popularity, larger budgets, and optimal runtime tend to generate higher revenue.

These factors likely play crucial roles in audience attraction and overall success at the box office.

Also, the original languages are naturally correlated to the production countries. The choice of original language often aligns with the cultural context of the production country. Cultural factors inherent in both language and country may influence the target audience and, consequently, the movie's revenue.

However, this paper doesn't see a significant relationship between revenue and release month.

3.3. Development of Random Forest

1. Establish a Random Forest regression model by training the dataset. This paper takes 70% of the data as a training data set and the rest 30% to test the model. The training data set is composed of the dependent variable revenue and the independent variables budget, popularity, runtime, release date, original language, and production countries.

2. Use the established Random Forest to compute the feature importance of each variable.

Calculate feature importance using the constructed Random Forest.

3. Apply the Random Forest regression model to both training and testing data to evaluate how the model performs.

3.3.1. Parameters of Random Forest

Table 1 presents the model parameters for the random forests

Table 1. Model Parameter of Random Forest

Parameter names	Parameters
Data split	0.7
Data Shuffling	Yes
Cross Validation	5
Node Split Criterion	mse
Max Feature Proportion in Split	None
Minimum Samples for Internal Node Split	2
Minimum Samples for Leaf Node	1
Minimum Weight of Samples in Leaf Node	0
Maximum Tree Depth	10
Maximum Number of Leaf Nodes	50
Threshold for Node Impurity in Split	0
Number of Decision Trees	100
Sampling with Replacement	True
K-fold Cross Validation	5

3.3.2. Cross-validation

To prevent overfitting on the training set due to improper dataset partitioning. This paper uses K-fold cross-validation to divide the dataset into k equally sized, mutually exclusive subsets. In each iteration, the model trains on the union of k-1, and tests on the remaining subset. This cross-validation algorithm will iterate for k times and the average of the estimation in each repetition is the final result of the algorithm [5].

Also, to ensure the effect of the cross-validation and the model running efficiency, this paper chooses the 5-fold cross-validation.

3.4. Development of LSE Linear Regression

The independent variable X in LSE linear regression comprises budget, popularity, and runtime, which are standardized quantitative variables. The dependent variable Y is a quantitative variable revenue being standardized.

The Results of model goodness-of-fit testing, the linear relationship between independent variables and the dependent variable, and so on.

3.5. Development of BP Neural Network

As the introduction of the method, the algorithms need to use the steepest descent method to keep modifying the weights and thresholds of the model to find the best-fit parameters of the network that lead to the lowest sum of the squared errors.

Similarly, the training data set is composed of the dependent variable revenue and the independent variables budget, popularity, runtime, release date, original language, and production countries.

In addition, the 5-fold cross-validation will also be applied. Table 2 presents the model parameters of neural network.

Table 2. Model Parameters of Neural Network

Parameter Name	Parameters
Data split	0.7
Data shuffling	True
k-th cross-validation	5
Activation function	identity
solver	lbfgs
Learning rate	0.1
L2 regularization	1
iterations	1000
The number of hidden neurons of the first layer	100

4. Results

4.1. The Result of the Random Forest

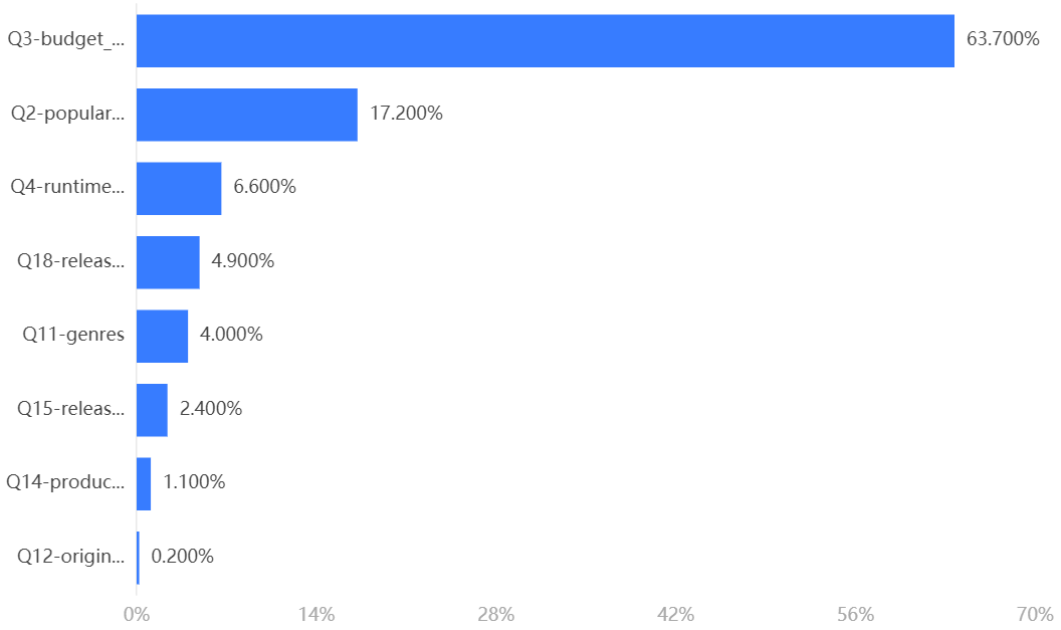

Figure 4. The eigenvalue of the model

As shown in figure 4, the result of the eigenvalue is similar to the correlation heat map done by Kendall's tau-b method, the budget is the most relevant independent variable and the next is the popularity. Table 3 shows the analysis result table of random variable.

Table 3. Analysis result table of Random Variable

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.001	0.025	0.013	52.572	0.881
Cross-validation set	0.002	0.049	0.019	60.208	0.541
Test set	0.002	0.039	0.018	64.673	0.738

4.2. The Evaluation Result of the LSE Linear Regression

Table 4 shows the analysis result table of linear regression.

y=-0.0 + 0.237* popularity + 0.023*Q8-runtime + 0.661*Q6-budget

Table 4. Analysis result table of Linear Regression

Analysis result of Linear Regression n=3000									
	Unstandardized coefficient		Standardized coefficient	t	P	VIF	R^2	adjustedR^2	F
	B	Standard error	Beta						
Constant	0	0.011	-	-0.014	0.989	-	0.608	0.608	F=1549.811 P=0.000***
popularity_	0.237	0.012	0.237	19.454	0.000***	1.131			
runtime	0.023	0.012	0.023	1.978	0.048**	1.067			
budget	0.661	0.012	0.661	53.239	0.000***	1.18			
Dependent variable： revenue									

4.3. The Evaluation Result of the BP Neural Network

Table 5 shows the analysis result table of neural network.

Table 5. Analysis result table of Neural Network

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.002	0.049	0.021	80.021	0.588
Cross-validation set	0.002	0.05	0.021	101.535	0.563
Test set	0.002	0.043	0.02	110.942	0.625

4.4. Comparison

This paper mainly uses the R square to evaluate the performance of the models. Overall, neural networks and linear regression have similar R squares around 0.6. The Random Forest performs best among these three models since its R square is 0.738. Also, it has the lowest Average Percentage Error. However, though the R square is of the model is relatively acceptable as a result of the high – uncertainty prediction model, the average percentage error is up to 60% so it might not eligible to be applied to real life business and need further improving in both modeling and the quantity and quality of the dataset.

5. Conclusion

In conclusion, the research shows that the budget, popularity, and runtime are the key correlations to predict the movie revenue before release.

While Random Forest excels, the achieved R^2 falls short of complete satisfaction. To enhance prediction accuracy, a strategic focus on feature selection and the development of more precise models is necessary. Careful consideration must be given to the potential influence of the popularity index from IMDB, warranting additional scrutiny and refinement.

Furthermore, the dataset used in this study is relatively out of date, which may impact the generalizability of the findings. It's important to recognize that the dynamics of the film industry evolve, necessitating a constant adaptation of models to changing trends.

To address these limitations and advance the research, future work should prioritize the following:

To improve these limitations, it should expand the dataset by incorporating more features that capture important aspects of movie performance. This can include variables related to evolving audience preferences, marketing strategies, and emerging trends in the film industry.

In addition, we can Increase the scale of the dataset to ensure a more comprehensive representation of diverse scenarios. A larger dataset allows for a more robust evaluation of model performance and enhances the reliability of predictions.

Also, given the dynamic nature of the film industry, ongoing data collection and integration are crucial. Continuously updating the dataset will ensure the models remain relevant and effective in predicting box office revenues.

By addressing these aspects in future research endeavors, the author will aim to refine predictive models, overcome current limitations, and contribute to a more accurate understanding of the factors influencing movie box office success.

References

[1] M. Ghiassi, David Lio, Brian Moon, Pre-production forecasting of movie revenues with a dynamic artificial neural network, Expert Systems with Applications, 42 (6): 3176 - 3193, 2015,

[2] H. Li. Statistical Learning Methods. Beijing: Tsinghua University Press. 7, 95 - 135, 2012.

[3] J. Cohen, P. Cohen, S. G. West, L. S. Aiken, Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates. 2003.

[4] Z. Zhou. Machine Learning. Beijing: Tsinghua University Press. 2016.

[5] A. Ramezan, Christopher, Timothy A. Warner, and Aaron E. Maxwell. "Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification", Remote Sensing, 11 (2): 185, 2019.