Predicting movie revenue from movie meta data

Pinzhi Huang [1,*], Siyu Chen [2], Ziyu Liu [3]

1 New York University, New York, New York, 10012, USA
2 Queen's University, Kingston, Ontario, K7L 3N6, Canada
3 Queen's University, Kingston, Ontario, K7L 3N6, Canada
**Corresponding author. Email: ph2239@nyu.edu*

ABSTRACT. *In this study, we apply machine learning techniques on the movie Meta data to predict box office movie revenue. This research use database collected from IMDb and processed with certain criteria. With the machine learning model, based on the parameter and database, this research is divided four parts including introduction to the research topic, steps to data processing, constructed the prediction model. The model is trained by movie meta data from the past movie and used to predict movie revenue from the data for a movie in early released. Research findings show 1) Model performance improved with additional features 2) Catboost was the best performing algorithm. The research shows with more features the prediction accuracy will increase.*

KEYWORDS: *box office revenue, machine learning, meta data, Gradient Boosting*

1. Introduction

Movies are a form of artistic expression in the cultural exchange between countries all over the world. Nowadays, movies gradually become a profitable commodity due to the film culture in today's society. Nonetheless, despite being a capital intensive industry, whether the success or profitability still mostly remains uncertain; thus, predicting movie revenue is now one of the most profitable research topics.

There is increasing numbers in use of machine learning as a new technology to predict how well a movie will perform in the future. Research showed Pruned Random Forest is able to find the future revenue for a movie in China [1]. Their studies show that the Pruned Random Forest have a better performance than other models and their model can do a prediction to show a range of revenue which a movie can go, which is useful for movie company to have the decision. Research study shows that using multiple models instead of single models will lead to more accurate predictions [2]. Also research by Nikhil specifies the box office revenue

prediction affected by the number of theaters on release, budget and first week-end revenue [3]. They predict movie revenue by combining k-means clustering and liner regression, showing that, in some cases it is possible to predict total revenue with accuracy better than 20%, while in many other cases are difficult to analyze due to lack of sample in database. Research using Support Vector Machine (SVM), combine with Neural Network and Natural Language Processing [4]. Techniques of them show different accuracy for pre-released features. Based on the research, they find the numbers of screens, Internet Movie Database (IMDb) voters and budget amounts are the most important feature which supports the most for predicting a movie's box office. With explored the Factorization Machines have the possibility to predict the success for the future of a movie by using rating data in IMDb for recent released movies by using the data from social media and combine it with the study they have [5].

In this research, this work use TMDB Box Office database and LightGBM and Catboost machine learning model to predict the movie revenue. This research mainly focuses on extracting data from the movies that have already been released to train the machine learning model for the further evaluation. Through this work it uses both the budget and popularity as features in machine learning model. The research will provide a model which will be able to use the features to predict estimate result of the total revenue for a pre-release movie.

2. Data

The research collected data for roughly 3001 movies including their rating, budget, popularity, release date, runtime and revenue. The dataset is extracted from IMDb (www.themoviedb.org), through use of their API. The research uses these movies and their features to try and predict their overall worldwide box office revenue. The research also added an additional dataset of 2888 movies from The Movie Database which provides us with additional features such as the popularity after the movies is released, total votes, and ratings. The research combined these two dataset for training the models.

Table 1.Features

Note: This table shows the features in this study used from IMDb database. The research uses these features both to generate graphs for analyzing their relationships and to train models for the machine learning and regression algorithm.

Feature	Description
IMDb_id	ID of a movie in IMDb database
Total votes	Total number of votes for the rating on one movie.
Rating	A measurement voted by users of how good or popular a movie is.
Budget	Budget of a movie in dollars. 0 values means unknown.
Popularity	Popularity of the movie in floating numbers.
Popularity_2	Popularity measurement of the movies after they released.
Release date	Release date of a movie in mm/dd/yy format.
Runtime	Total runtime of a movie in minutes.
Revenue	Total revenue earned by a movie in dollars.

Table 2.Summary Statistics

Note: This table illustrates summary statistics for the features used in this analysis.

Feature	Average	Minimum	Maximum	Median	Standard deviation
Total votes	993.94	1	18931	18	1744.189
Rating	6.36	1	9	6.2	1.143
Budget	22531334.1	0	380000000	8000000	37026086.4
Popularity	8.46	0.0001	294.34	7.374	12.104
Popularity_2	8.029	0.6	45.15	2.349	5.211
Runtime	107.86	0	338	105	21.887

2.1Criteria

Out of the 3000 movies, this research filtered out movies that had a budget less than $1,000, which resulted in 2800 movies for the analysis. Pre-processing resulted in removing 11 movies inform the data due to a runtime value of 0. After cleaning the data, this research has 2158 movies use for the training process.

2.2Release date

As the number of years grows, so does the number of movie releases. Total movie box-office revenue has an upward trend and fluctuated in the period of almost one century (1920-2016). It is worth to notice that the total revenue has a sudden growth between 1972 and 1976 and reached its peak in 1974. There is a seasonality effect where movie revenue is higher in the months of June and December but clearly lower in the other months.

2.3Rating count

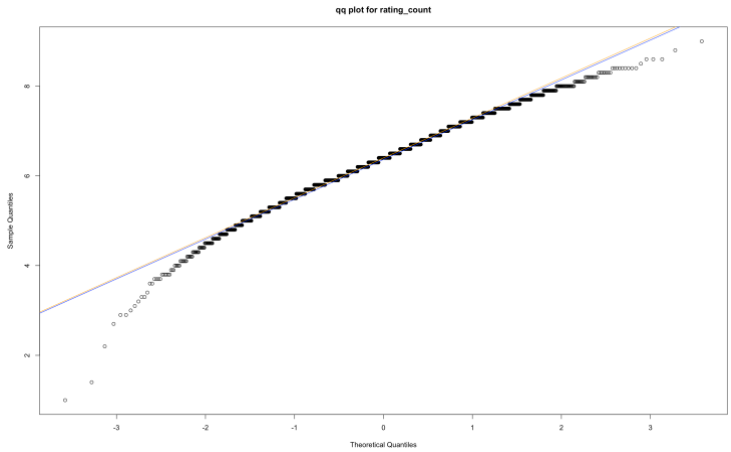

Figure1.Q-Q Plots for rating count

Note: This figure shows a Quantile-Quantile Plot for rating count feature. A q-q plot is a plot of the comparism of two different dataset in order to evaluate the relation for both dataset.

QQ plots compare the distribution of two quantile. (Figure 1) In this study, the horizontal axis shows quantiles from a theoretical normal distribution. It is then compared with a set of data on the vertical axis, which is the distribution of the samples(rating counts) The line fits the data points well, which implies the rating count almost follows the typical pattern of the normal distribution. This distribution

is symmetric with most data concentrated at the mean and medium and less data on both sides.

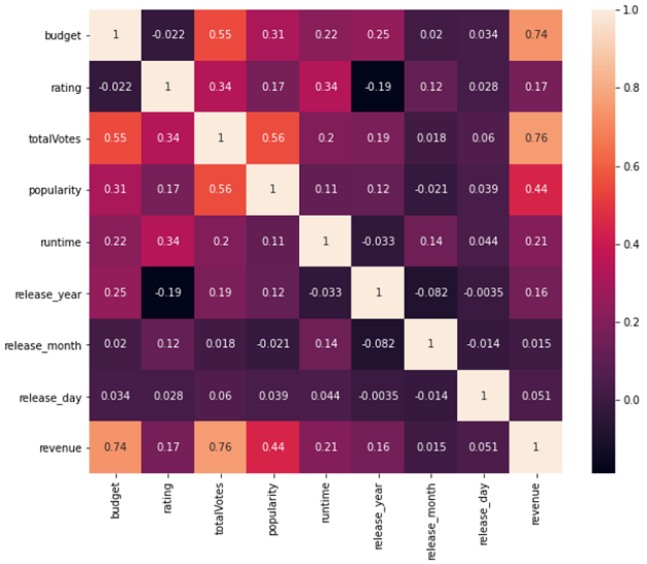

Figure2. Correlation of Features

Note: This is a heat map for the correlation of features used in this analysis. The bar on the right with different colors represents different degrees of the correlation.

The above heat map displays the degree of the correlation between different factors. The degree of correlation is represented by the coefficient between 0.0 and 1.0. Movie revenue is strongly correlated to the budget with a correlation of 0.74. This means movies with higher investment are more likely to receive higher rewards, which refers to the box office. Also, movie revenue has a strong relationship with total votes with a correlation coefficient of 0.76. This implies movie revenue is influenced by the social sentiment. With higher ratings, movies are more likely to generate higher profits. Popularity is also an important proxy of high movie revenue with a correlation coefficient of 0.44.

3. Methodology

The Proposed methodology is illustrated in Figure 3. It contains following steps:

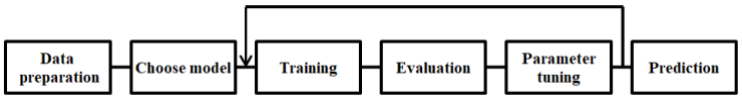

Figure3.methodology

Note: This figure shows a flowchart for the methodology in this research. The flowchart shows the steps as boxes of this research method and the order we perform by connecting the boxes with arrows. It contains 6 steps in total to explain the research approach to solve the problem.

3.1Data preparation

The dataset is provided by IMDb collected from the TheMovieDB.ORG (TMDB) Open API. The research imports the data, analyze the dataset and then clean data. The data has irrelevant features and missing values. To solve this problem, only choose the features that have a strong correlation with the revenue. For missing values we join the IMDb dataset with TMDB to fill in the missing values , and apply any filtering to remove erroneous information.

3.2Choose model

The research is to train 3 models, Gradient Tree Boosting (GBT), Light GBM and Catboost with the same training set for performance comparisons. The models chose are discussed below.

GBDT

Gradient Boosting Decision Trees (GBDT) is a generalization of boosting to arbitrary differentiable loss functions. GBDT is well performed off-the-shelf procedure that is useful for solving both regression and classification problems in a variety of areas in doing predictions for numbers or ranking [6]. Figure 4 illustrates the training process of GBDT Algorithm. In the process of learning and weighting, it calculates the residuals by comparing the difference between actual data (previous decision tree) and predicted data (current decision tree), then the new decision tree will be constructed by the residual values and each classifier continues training on the formed residuals through each iteration by the same process in order to reduce the error and combine all constructed trees to obtain a more accurate ensemble model.

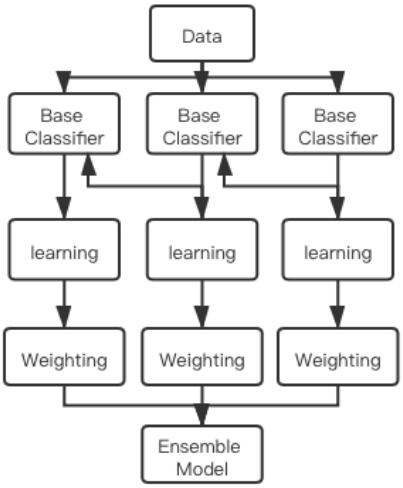

Figure4. GBDT Algorithm Training Process

Note: Figure4 shows a flowchart for Gradient boosting Algorithm, it contains base classifier, learning and weighting process, the goal is to produce a prediction model and give a prediction result.

LightGBM

Another machine learning based algorithm that the research applies for the movie box-office revenue forecasting is LightGBM (Light Gradient Boosted Machine), which is a highly efficient gradient boosting decision tree bases on histogram algorithm. Inventor emphasized on the problem of the high computational complexity due to the massive features and instances when predicting with GBDT [7]. They developed LightGBM, which is the novel GBDT algorithm based on Gradient-based One-side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to handle the large of data instances. LightGBM performs the parallel learning in both features and data, the parallel algorithm assists to form the decision tree by finding the best split of data information, and it largely improved the efficiency of training process. For the regular decision tree based algorithms, they usually follows the level-wise tree growth principle, while the LightGBM is a special one with leaf-wise tree growth, the tree will only choose the leaf with max delta loss and keep growing on that leaf node.The result of the experiment clearly shows that the LightGBM has better trade-off in training efficiency and memory usage when comparing with GBDT algorithm. LightGBM is a gradient boosting algorithm that uses tree based learning algorithms and designed to be distributed and efficient with

the following advantages:

- Higher performance in training process.
- Less memory space needed.
- More accurate compare to other similar model.
- Able to install GPU learning to speed up.
- Capacity to process meta data at one time.

Catboost

Catboost algorithm is a novel machine learning algorithm developed by Yandex, a famous Russian searching engine. Research applied ordered boosting with ordered TS (target statistics) to solve the problem of prediction shift due to the defective currently existing gradient boosting algorithms [8].By comparison, the ordered boosting mode of Catboost performs better than XGBoost and LightGBM, especially in the accuracy performance. CatBoost combines the categorical features and gradient boosting, also based on the decision tree library. It is a powerful machine-learning technique that achieves state-of-the-art results in a variety of practical tasks. Catboost has algorithm for transforming the categorical data into numerical data and takes advantage of dealing with them during training as opposed to preprocessing time. Another advantage of the algorithm is that it uses a new schema for calculating leaf values when selecting the tree structure, which helps to reduce the overfitting. Steps for the Catboost training process are to get the residuals from each data point and apply the trained model on all remaining data. The research is to train the model with the residuals of each data point, these steps are repeated until the algorithm converges.

4. Results

4.1Training and parameter tuning

The research split the dataset into 80% train and 20% test. After training the model and evaluate the result to adjust the model parameters through an exhaustive parametric grid search. Grid search cross-validation is used to train a machine learning models for each combination of model parameters to determine the optimal combination of parameters which yields the highest model performance. Model performance is measured through the R^2 regression score as the performance metrics in this research to help us to choose the best model of all time.

4.2Prediction

Using the holdout data set (20%) The research is to make predictions for all three models and compare r-squared (R^2) score and other performance measures from each model (Table 3).

Table 3.Performance for each model

Note: This form draw a comparison between three models, and shows the R^2 score as well as running time they have on making predictions. The training and tuning time is in seconds (s).

	GBDT	LightGBM	Catboost
R^2 score(without additional feature)	0.64	0.53	0.64
R^2 score	0.76	0.73	0.78
Training time	34 secs	15 secs	8 secs
Parameter Tuning Time(for 27 fits,500 iteration）	315secs	168secs	238secs

Catboost yielded the best performance in predicting movie revenue. All the models not perform well without the additional feature which included rating, total vote and popularity. It proves that the number of the features is essential for a machine learning model to have a well performance in prediction problem. As it shows LightGBM gain the most progress after the feature is added. It proved that LightGBM is using the less important features more effective than others. In the study can see CatBoost comes out as the maximum accuracy on test set and minimum training time. GBDT generally works well and with less parameters set up compare to others, it's the easiest model to set up. For LightGBM got the minimum parameter training time, in terms of accuracy it reaches a slightly lower score as others.

5. Conclusion

The research successfully builds up machine learning models to help us to predict the movie revenue. We are able to compare performance differences between these models and result show Catboost is the best model in this dataset. However, the result may vary if the dataset has more features or numbers of movies to analyse. For further study, look for more factors which will also affect the revenue and the study are certain that with larger dataset this research will be able to improve the reliability and accuracy in the prediction model.

References

[1]. Z. Guo, X. Zhang, and Y. Hou, “Predicting Box Office Receipts of Movies with Pruned Random Forest,” in *Neural Information Processing*, Cham, 2015, pp. 55–62.

[2]. G. He and S. Lee, “Multi-model or Single Model? A Study of Movie Box-Office Revenue Prediction,” in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, Oct. 2015, pp. 321–325.

[3]. Nikhil Apte, Mats Forssell, Anahita Sidhwa “Predicting Movie Revenue“http://cs229.stanford.edu/proj2011/ApteForssellSidhwa-PredictingMovieRevenue.pdf

[4]. N. Quader, Md. O. Gani, D. Chaki, and Md. H. Ali, “A machine learning approach to predict movie box-office success,” in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, Dec. 2017, pp. 1–7.

[5]. B. Çizmeci and Ş. G. Ögüdücü, "Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media," *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, 2018, pp. 173-178.

[6]. J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[7]. G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” p. 9.

[8]. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” p. 11.