Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/29/20230742

Prediction of movies popularity in supervised learning techniques

Yizhen Zhang[1,3] and Zejun Bai[2]

[1]School of Engineering, Columbia University, New York City, New York 10025, United States
[2]School of Electrical Engineering, Sichuan University, Shuangliu District, Chengdu, Sichuan 610065, China

[3]zyizhen1998@outlook.com

Abstract. When movies industry gradually become heavy capital, the prediction of movies' popularity as well as their commercial potentiality based on historical data has become a popular research topic in the field of data analysis using machine learning models. In this paper, researchers trained three supervised machine learning models (Random Forest, Naive Bayesian Model and Support Vector Regression) using IMBD dataset to predict a movie's popularity. This research has two outcomes: (1) the Random Forest Model has the highest accuracy rate; (2) the number of Oscar-winner included in both cast and crew is most positively related to a movie's popularity.

Keywords: predictive analysis, unsupervised machine learning, random forest, Naive Bayesian Model, Support Vector Regression.

1. Introduction

Nowadays, movies are popular market products consumed within a cosmopolitan commercial milieu. To provide panoramic view on the current movie market, this paper would begin with a generic sketch of the interrelated history of development of global commerce and movie industry. The first ballooned expansion of commerce started in the end of Mid-Age Europe, when self-sufficient manors were gradually developed into mercantilely active town, which later became populated cities functioning as trading centers. Both the scale and frequency of global commerce augmented as human civilization evolved through the Scientific Revolution, the Age of Navigation, Industrial Revolutions and Enlightenment, finally bloomed after the newly founded America gradually became a worldwide fiscal center in the 19th century. During this period of swift economic growth, the entertainment industry, previously mostly inclusive to upper classes, quickly expanded with a trend towards universality in that a general improvement on living condition ascended both the needs and the affordability for commonalty towards entertainment products.

On the eve of the 20th century, rendering sensual experiences in view of projected cinematographic motion pictures, the film emerged as an innovative entertainment product in European and soon gained worldwide popularity within fifty years. Different for the experimental embryo created during their nascent stage, film products nowadays become professionally exquisite and are largely produced by a

DOI: 10.54254/2755-2721/29/20230742

capital-heavy industry, which includes writing, cast and crew selection, actual shotting, editing, commercial propaganda, screening, and launch events. Etc. The efficacy of a mature industrial chain and the perversity of film products ensured the consistently high production rate of new films. In the past 2020, over 203 thousand movies successfully go on screen. Though earning the least revenues since 2016 due to the impact of coronavirus, the entire global theatrical and home/mobile entertainment market in 2020 still totaled $80.8 billion and showed healthy projection for the future based on the growthy trend of worldwide office revenue since 2016 [1].

The rapid expansion of movie market is tightly correlated to the recent application of data science within the industry. From determination of production cost to design of market campaign, numerous factors are at play in filmmaking process professionals employed Data-scientistic techniques on onetime data to make decisions ensuring commercial success. Since the revenue of a film is strongly influenced by its popularity, the popularity figure predicted by data-scientistic model would be useful to foretell a film's potentiality to achieve commercial success before its production. This paper will use machine learning to tackle historical values extracted from TMDB website and established predictive analysis to foretell a film's popularity before its release to audiences. Data variables used to predict popularity included budget, cast and crew selection, and plot overview.

2. Literature review

During the process of model-establishing of rating prediction system, many previous studies used basic machine learning model such as linear regression, support vector regression (SVR), random forest, Naïve Bayesian Model (NBM), convolutional neural network (CNN), k-Nearest Neighbor(k-NN).

Linear regression is perhaps one of the most well-known and well understood algorithms in the field of machine learning. In statistics, linear regression is a linear approach to modelling the relationship between the explanatory variables (independent variables) from input database and the scalar response (dependent variables). Ping-Yu Hsu, Yuan-Hong Shen, and Xiang-An Xie had concluded that in the past 25 years, marketing scholars mainly used multiple linear regression to form a forecast popularity model. Hsu, Shen and Xie established their multiple linear regression model using user rating as dependent variables and other variables (predictor variables) as independent variables, when the latter was selected piecemeal using stepwise regression technique [2]. Different from Ping, Yuan and Xiang using historical data from IMDb, Oghina, Breuss, Tsagkias and De Rijke took Twitter content as textual data for predictive text mining. This choice of independent variables over the regression model postpones the application of predictive model till the movie has been released [3]. Schmit and Wubben used similar methodology to predict ratings on IMDb [4].

Schmit and Wubben also used support vector regression (SVR) in the tasks of data regression. SVR is also popularly and widely used for regression problems in machine learning. As in regression, support vector regression (SVR) is marked by its use of kernels, sparse solution, and VC control of the margin and the number of support vectors [5].

Besides regression model, other algorithms such as ensemble is also frequently used to tackle data scientistic problems. One of the most renown ensemble algorism, Random Forest combines several independent weak classifiers and then makes the final decision by voting or taking the mean. In the research field of movies' popularity, Hemraj Verma, Garima established and compared various prediction models for Bollywood Movie Success and random forest renders calculable performances among all [6]. K. Pradeep, C. R. Tintu Rosmin, Sherly Susana Durom, G. Anisha used Decision Tree Algorithms for Accurate Prediction of Movie Rating [7]. Another classification algorithm is Naive Bayes Classifier that based on Bayes' theorem and the assumption of independence of characteristic conditions. S. Indhu, S. Lavanya did a survey on the unsupervised joint topic modelling approach in Bayesian model [8]. V. S. Reddy, D. Somayajulu, A. Dani used the classification of movie reviews using complemented Naive Bayesian Classifier [9].

Abarja, R. A. and Antoni W. used convolutional neural network (CNN), "a more complicated machine learning technique contained in the class of artificial neural network, to work on the same project. Always used in deep learning, CNN is a more complicated machine learning technique

DOI: 10.54254/2755-2721/29/20230742

contained in the class of artificial neural networks. Abarja and Antoni's study proved that one-dimensional CNN had promising result as model to predict movie rating based on the tabular dataset. Therefore, CNN could also be used as model for small textual dataset.

To the end, the last machine-learning method included in the literature review is the k-Nearest Neighbor(k-NN)- a standard machine-learning method that has been extended to large-scale data mining efforts. In related scholarship, S. Kabinsingha, S. Chindasorn and C. Chantrapornchai applied data mining to the movie classification and run experiments that have about 80%-88% precision for all the tested rating [10]. O. B. Fikir, İ. O. Yaz and T. Özyer utilized matrix value factorization for predicting rating i by user j with the submatrix as k-most similar items specific to user i for all users who rated them all [11].

3. Data

In total, four datasets are included in this research: TMDb 5000 Movies, TMDb 5000 Credits, World Top Ten Movie Company, Oscar Winners. TMDb 5000 Movies dataset, collected through TMDb database, a popular user editable database, described 4800 distinct movies using 4 labels including cast and crew. TMDb 5000 Credits evaluated the same movies with other 23 labels including genres, revenues, budgets, and popularity. World Top Ten Movie Company included the top ten movie companies (2020) in the world. Oscar Winners recorded all the Oscar winners in history.

4. Data-preprocessing

Through the data-preprocessing progress, this research established a new dataset (Data) with seven new features (budget level, popularity level, revenue level, runtime level vote average, Oscar cast, Oscar director, production company) from the original data sets through data normalization and data cleaning. Table 1 illustrates the specific data-preprocessing progress, including the seven features selected and their data-cleansing process in details. The head ten lines from the new dataset after preprocessing are presented in Table 2.

Table 1. Preprocessing progress.

Production Company	**The number of world top-ten companies participated in production**
Budget Level	a level of 1 to 8 after a categorization of numerical data in "budget" [Tmdb_5000_movies.csv"] based on shape of the dataset's distribution
Popularity Level	A similar level of 1 to 8 after a categorization of numerical data under key "popularity" [Tmdb_5000_movies.csv"]
Runtime level	A similar level of 1 to 8 after a categorization of numerical data under key "runtime" [Tmdb_5000_movies.csv"]
Oscar Director	The number of Oscar-winner directors participated in the movie production
Oscar Cast	The number of Oscar-winner cast participated in the movie production
Vote Average	Data under key "vote average" from [Tmdb_5000_movies.csv]
Revenue Level	A similar level of 1 to 8 after a categorization of numerical data under key "revenue" [Tmdb_5000_movies.csv"]

Table 2. New dataset.

	Production_companies	**Buget_level**	**Popularity_level**	**Runtime_level**	**Oscar_level**	**Oscar_cast**	**Vote_average**	**Revenue_level**
0	0.0	1.0	8.0	1.0	1	0	7.2	1.0
1	1.0	1.0	8.0	1.0	0	0	6.9	1.0
2	1.0	1.0	8.0	1.0	1	0	6.3	1.0
3	1.0	1.0	8.0	1.0	0	0	7.6	1.0

DOI: 10.54254/2755-2721/29/20230742

Table 2. (continued).

4	1.0	1.0	7.0	1.0	0	0	6.1	1.0
5	1.0	1.0	8.0	1.0	0	0	5.9	1.0
6	1.0	1.0	8.0	1.0	0	0	7.4	1.0
7	1.0	1.0	8.0	1.0	0	0	7.3	1.0
8	1.0	1.0	8.0	1,0	0	0	7.4	1.0
9	2.0	1.0	8.0	1.0	0	0	5.7	1.0
10	1.0	1.0	8.0	1.0	0	0	5.4	1.0

5. Methodology

In the design of this research, data under the key "popularity level" in the new dataset is established as an indicator of a movie's popularity, the predicting value. The numerical data, all integers, in this key ranges from one to eight and the size of integers is positively related to the movie's level of popularity.

Three machine learning models are chosen for this topic: Random Forest, Naive Bayesian Model (NBM) and Support Vector Regression (SVR), the first two being classification models and the last regression model. A ratio of 2:8 was used to split the training set and testing set.

Three second-level parameters (accuracy, precision, and sensitivity) from confusion matrix are used as evaluation criteria to compare the performances among three models. A general evaluation on model's performance will be given upon a comprehensive consideration among all three parameters. Table 3 shows the raw data generated by the confusion matrix for seven features respectively. Table 4 presents the numerical evaluation given by the confusion matrix. Figure 1 converts data in Table 4 into linear representation. In summary, a flow chart that sums up the methodology conducted in this research is presented in Figure 2.

Table 3. Result of confusion matrix.

	Confusion matrix	**Variables(keys)**	**Numbers of Correct Prediction**
1	[9 6 17 30 37 14 0 0 0 0]	Production Companies	113
2	[0 0 7 21 38 26 9 6 0 0]	Budget Level	107
3	[0 0 4 13 29 37 26 6 0 0]	Popularity Level	115
4	[0 0 0 13 36 37 42 5 0 0]	Runtime Level	133
5	[0 0 0 6 20 42 44 15 0 0]	Oscar_ Director	127
6	[0 0 0 3 16 31 47 24 1 0]	Oscar Cast	122
7	[0 0 0 3 9 26 49 35 3 0]	Vote Average	125
8	[0 0 0 2 3 10 31 59 13 1]	Revenue Level	119

Table 4. Numerical representation of models' performances.

	Random Forest	**NBM**	**SVR**
Accuracy	0.538833	0.506752	0.573414
Precision	0.528881	0.486462	0.15343
Sensitivity	0.524219	0.482944	0.095951

DOI: 10.54254/2755-2721/29/20230742

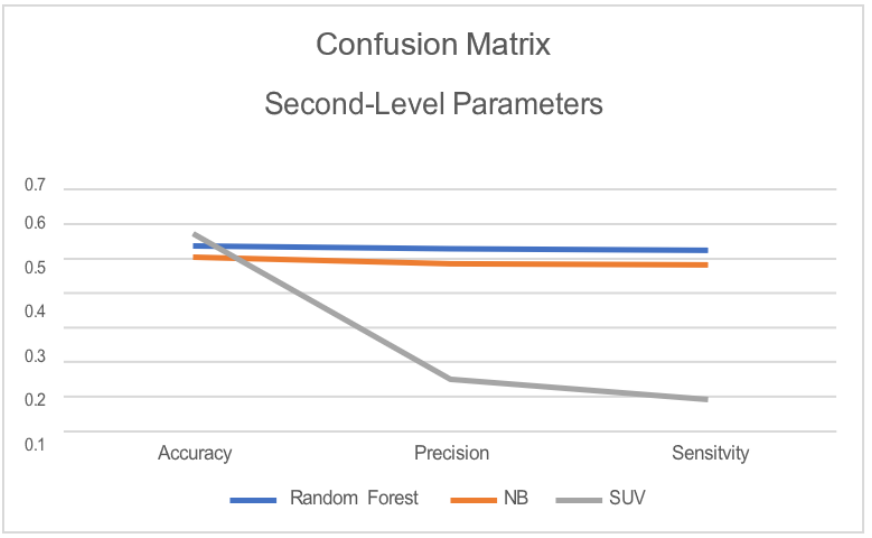

Figure 1. Linear representation of models' performances.

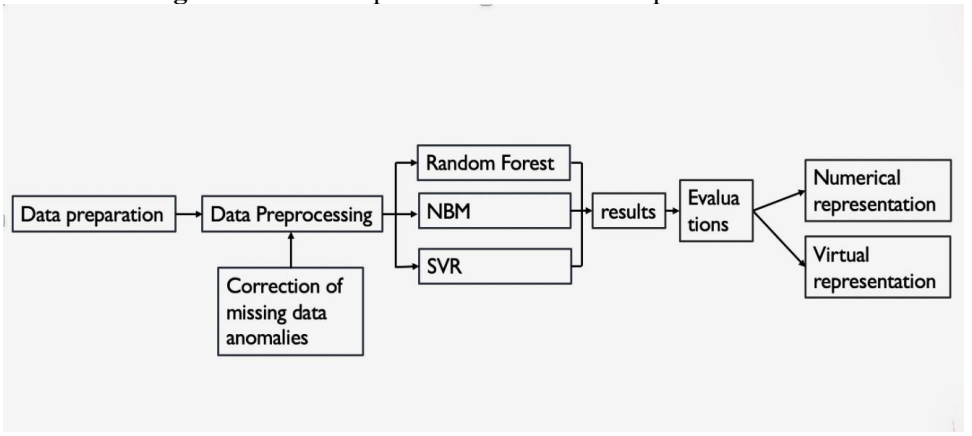

Figure 2. Methodology flow chart.

6. Conclusions

After a comparison on the performances for each model, Random Forest Model was evaluated as rendering the best performances from its consistent efficacy on accuracy (0.54), precision (0.53) and sensitivity (0.53). NBM showed similar stability in all three parameters, but averagely scored 0.02 below the random forest model. Though SVR had the highest accuracy score (0.57), its poor performance on the rest two parameters stemmed it from being rated as an effective predictive model for movies' popularity. Then, this research evaluated the variables' contributions towards the prediction using random forest model, which has been proved as the most reliable as above. Credited to an application of confusion matrix, the evaluation result listed the top two influential variables to a movie's popularity: runtime level and Oscar Director.

Therefore, based on the IMBD dataset, this research concludes that two primary factors that are positive-linearly related to a movie's popularity are its overall runtime and its director. The commercial success of a movie is usually guaranteed by a common runtime and the direction of a past Oscar winner. This research briefly summarizes the commercial code to produce a popular movie and provides a scientific proof to investors who are interested in movie industry.

References

[1] Abarja, R. A. (2020) Movie rating prediction using convolutional neural network based on historical values. International Journal of Emerging Trends in Engineering Research 8, 5: 2156–2164.

[2] P. Y, Hsu, Y. H. Shen and X.A. Xie (2014) Predicting movies user ratings with Imdb attributes. In: International Conference on Rough Sets and Knowledge Technology. Cham.
[3] A. Oghina, M. Breuss, M. Tsagkias and M. de Rijke. (2012) Predicting IMDB movie ratings using social media. In: Advances in Information Retrieval, Barcelona, 2012.
[4] W, Schmit and S. Wubben (2015) Predicting ratings for new movie releases from Twitter content. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Lisboa.
[5] Abidi, S. M, Y. Xu, Ni J, X. Wang and W. Zhang. (2020) Popularity prediction of movies: From statistical modeling to machine learning techniques. Multimedia Tools and Applications, 79: 47-48.
[6] Verma, H., & Verma, G. (2019). Prediction model for Bollywood Movie Success: A Comparative Analysis of performance of supervised machine learning algorithms. The Review of Socionetwork Strategies, 14(1), 1-17. doi:10.1007/s12626-019-00040-6
[7] Pradeep, K., TintuRosmin, C. R., Durom, S. S., & Anisha, G. S. (2020). Decision tree algorithms for accurate prediction of movie rating. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). doi:10.1109/iccmc48092.2020.iccmc-000158
[8] Indu, S., & Lavanya, S. R. (2018). A Survey on Unsupervised Joint Topic Modeling Approach in Bayesian Model. Biometrics and Bioinformatics, 10(4), 66-69.
[9] Reddy. V., S. R., Somayajulu, D. V., & Dani, A. R. (2011). Classification of movie reviews using complemented naive bayesian classifier. International Journal of Intelligent Computing Research, 2(3), 148-153. doi:10.20533/ijicr.2042.4655.2011.0019
[10] SivaSantoshReddy, A., Kasat, P., & Jain, A. (2012). Box-office opening prediction of movies based on hype analysis through Data Mining. International Journal of Computer Applications, 56(1), 1-5.
[11] Fikir, Ozan & Yaz, İlker & Özyer, Tansel. (2013). Movie Rating Prediction with Matrix Factorization Algorithm. 10.1007/978-3-7091-1346-2_28.