Movie Rating Prediction Using Machine Learning

Grigor Keropyan*
Department of Mathematics and Mechanics
Yerevan State University
goqorkeropyan@gmail.com

Abstract

Movie rating prediction based on information available prior to theatrical release is important in order to understand how successful will be movie. This paper describes various Machine Learning methods to predict movie rating.

1 Introduction

Movie rating prediction is becoming a popular problem and various methods have been suggested. In this paper we will introduce Machine Learning (ML) methods to predict movie rating based on the data available prior to the theatrical release. We used IMDB open-source data such as posters and run-time of movies.

Although, it seems impossible to predict the movie rating based on the information available before theatrical release, we will see that ML methods predict them quite accurately. One of the explanation could be that people may like movies if their preferred actor is in the crew or films are produced by their favorite producers.

One of the goals of this work is provide a tool which can help producers to promote their films to be successful. Another one is that this work provide a good recommendation for people predicting their IMDB score.

2 Related Work

This work is based on the Stanford students paper [1] where they predict IMDB score of movies. In [1] authors used ML based methods to predict IMDB score. Another research which uses Bayesian approach [2]. This paper is based on [1] and uses new ML algorithms such as lightgbm to reach the better accuracy than [1]. Our preproccessing is almost the same as [1] and we have about 20K examples. We have shuffled the dataset and splited into training and test test correspondingly 80 and 20 precents. In the algorithm we have done K-Folds cross validation and after tuning the hyperparamters we have achieved better accuracy than [1].

3 Dataset and Features

We have used open source data, "Movie Genre from its Poster Dataset" [3] and "The Movie Dataset" [4]. The dataset is preprocessed like [1], taking only movies which original language is english and which produced after 1980. The features are divided into 3 groups: text (synopses), images (posters) and others (runtime, genre, director, actors). From poster we have extracted some features: such as

*https://github.com/grigor97

1st ML project (ASDS 2020), Armenia.

Table 1: Dataset by Groups

Data	Type	Dimension	Example
Actors	Categorical	1671	Tom Hanks
poster	Numerical	13	Number of faces = 0
Director	Categorical	491	Howard Deutch
Runtime	Numerical	1	121 (minutes)
Genre	Categorical	23	Documentation
Synopses	Categorical	3712	Once

number of peoples in the poster, mean and standard deviation of blue, green, red, hue, saturation, brightness as in [5]. For summaries we have used count vectorizers and only left the word which appeared more than 20 times. Directors are have been made one hot and actors count vectorizers where we have only left top three actors in each movie. The example of result is shown in Table 1.

After filtering the english movies and released after 1980 it remained 18850 movies. Which we have splited in two ways. First: like [1] implementation, when it is separated train, validation and test sets correspondingly 70, 15, 15 percents. Second: which gave a better result it is separated into train and test sets correspondingly 80 and 20 percents.

4 Methods

In this chapter we will speak about the models which have been tried for the dataset.

4.1 Linear Regression

Linear Regression (LR) is on e of the widely used ML algorithms. It is simple and performs well in many tasks. LR discovers linear relationship between input and target data. In many real world applications target is actually linearly dependent of the input data. So, in this problems LR is a good choice.

In order to avoid overfitting we have also used Ridge Regressor. For this dataset latter one performed better than the former one. The reason could be that in that dataset exist outliers which deviate LR to perform well. Using the regularization model performed sagnificantly better.

4.2 Decision Tree and Random Forest

Decision Trees and Random Forests are commonly used interpretable models. Decision Tree model in each point of tree divides some feature in order to maximize targeted loss function. Random Forest is an assemble of Decision Trees which could perform better in complex data case.

As we will see in the next section Random Forest performed better than LR. The reason is that we have a a lot of categorical features which is more easy for Random Forest to separate from each other.

4.3 XGBoost

XGBoost is a new model compared with former ones. This model also an assemble of trees, but it adds regularization term to the loss function and uses another technique to find an optimal solution.

XGBoost also performed well on the dataset and it is better than LR models. Actually, Without fine tuning it performed a little worse than Random Forest.

4.4 Support Vector Regressor

Support Vector Machine is widely used ML algorithm which performs well in many problems. It defines a margin between points and only points beyond the margin affect on the loss. SVR instead, defines a margin of epsilon-distance and points inside the boundaries of regions do not affecting to loss.

4.5 Lightgbm

Lightgbm is based on XGBoost and it performs better than XGBoost in many tasks. It also very fast algorithm. We have used this model and after fine tuning it produces the best result. It outperforms to the results in [1] sagnificantly.

5 Experiments and Results

6 Discussion

askdnvkadsnfvkdafnk

7 Conclusion and Future Work

dsfavdfvd

References

[1] Yichen Yang et al., *"Predicting Movie Ratings with Multimodal Data"*, `http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26260680.pdf`.

[2] Y. J. Lim and Y. W. Teh, *"Variational bayesian approach to movie rating prediction"* Proceedings of KDD Cup and Workshop, vol. 7, 2007, pp. 15–21.

[3] KaggleInc, *"Movie genre from its poster,"* `https://www.kaggle.com/neha1703/movie-genre-from-its-poster` .

[4] Kaggle, *"The movies dataset,"* `https://www.kaggle.com/rounakbanik/the-movies-dataset` .

[5] F. B. Moghaddam, M. Elahi, R. Hosseini, C. Trattner, and M. Tkalčič, *"Predicting movie popularity and ratings with visual features,"* in 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). IEEE, 2019, pp. 1–6.