TMDB BOX OFFICE PREDICTION

SHUKAI WANG

Abstract. With the development and maturity of machine learning, more and more problems are associated with machine learning. This paper introduces the application of machine learning in film box office.

With the improvement of people's living standard, more and more people relax themselves through movies in their spare time. It also attracts a lot of companies to invest in it, but it also takes a lot of risks. In this paper, machine learning algorithm is used to predict the box office of a movie according to some existing film information to determine the possible income of the film, and then to determine the risk of investment in the film. This article will show the relationship between various factors and box office from many aspects, and then determine the factor training model closely related to it, so as to improve the accuracy of Prediction.

Contents

Date: 2020-10-13.

2020 *Mathematics Subject Classification.* Artificial Intelligence.

Key words and phrases. Machine Learning, Data Mining, Box Office Prediction.

1. Introduction

In 2018, film revenue has increased significantly, and the film industry is more popular than ever. What kind of movies make high box office receipts. In the process of preparation and shooting, whether the budget, the number of directors and actors have a great impact. Whether the publicity and preview of later films will affect the final box office income of films.

Data Analysis aims to show the relationship between attributes and box office revenue according to the data provided. And further integration of data, delete irrelevant data, unified data values and so on.

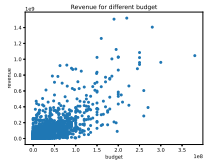
Models and Forecasts aims to use the integrated data to train the relevant models, improve the accuracy, and make the box office revenue forecast for some of the given data.

In this paper, we train the random forest model with the integrated data, and use the model to predict the box office of movies.

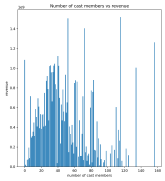
2. Preliminaries

In the early stage, the existing data were visually analyzed.

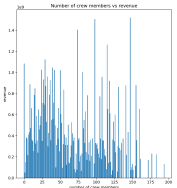
In the pre production budget, the number of actors and the number of directors and other crew members on the impact of box office.



Budget

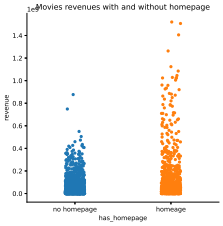


Cast

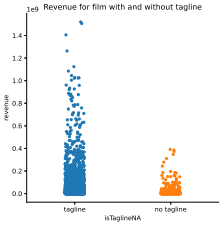


Crew

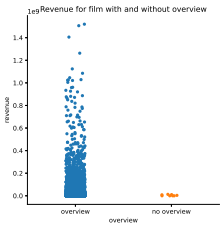
In the late stage of film production, the influence of home page, film introduction, film tagline on the box office of films is also discussed.



Homepage



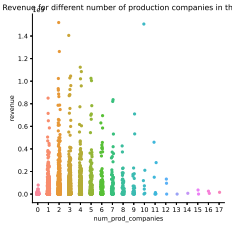
Tagline



Overview

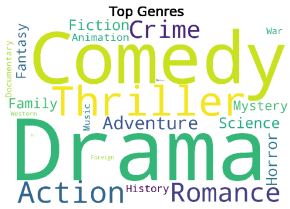
Some other influences, such as film series and the number of film companies participating in the investment.

File Series



File Company

The word cloud shows the movie types and languages that are liked by the audience.



Genre

Language

3. Method

After the preparation work is completed, the data will be further processed. Delete data not related to the box office, For example: imdb_id, orginal_title, poster_path, status. Normalize some data,For example: has_homepage, collection, overview, isTaglineNA, Keywords_count etc

After the completion of data processing, the model was established using random forest algorithm.

Random Forest Algorithm is an ensemble technique that combines multiple decision trees. Other advantages of random forests are that they are less sensitive to outliers in the dataset and don't require much parameter tuning.

4. Conclusions

The investment in the early stage and publicity in the later stage have an impact on the box office.

In the early stage, the number of actors and crew should be moderate, not the more the better.

The prediction accuracy can be further improved, such as using XGBoost.

Acknowledgment

Thanks for the learning opportunities provided by Tulip, I grew up rapidly in this period of time. Thanks to my tutor for answering questions and puzzles during my study, which is my rapid progress. At the same time, I should also like to thank my senior brother, sister and classmates for their help. When I am not successful in my study, I will lend a helping hand to help me solve problems.

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: 1450479286@qq.com