Box Office Prophet: a Machine Learning Study

Anna Deng, annadeng2020@u.northwestern.edu
Hayden Udelson, hudelson@u.northwestern.edu
EECS 349 Machine Learning, Northwestern University

Synopsis

The global film industry is a major force not only in culture but economics. Global box office revenue is expected to reach $50 billion by 2020 as it continues to grow. In the United States, the film industry is expected to generate $35.3 billion in revenue in 2019. A successful film can not only generate large amounts of revenue, but shape cultural dialogue long after its run in theaters has expired.

Our goal is to apply machine learning techniques to predict whether or not a film will be successful based on a number of attributes of that film. For the sake of this project, we interpreted success as whether or not a film earns more money in the box office than it cost to produce. A film is an artistic endeavor, with a large overhead cost and a large degree of uncertainty as to its profitability. Further, the success of a film is highly uncertain. Roughly half of movies produced by Hollywood turn a profit in the box office, and this is true of our data set as well. Of 4,157 films in our dataset, only 55.7% were profitable. By creating a machine learning model to predict a film's success, we can shed some light onto what makes a film a hit versus a flop.

Data

We began with a dataset from Kaggle.com, containing data on 5,044 films. This data was scraped from IMDb, the Internet Movie Database, in November 2016. The dataset has data on the following 28 attributes for each film:

- Color or black and white
- Director name
- Number of critic reviews
- Duration
- Likes on director's Facebook page
- Gross (domestic box office revenue)
- Genres
- Title
- Number of user votes
- Total likes on cast's Facebook pages
- Number of faces in movie poster
- Plot keywords
- IMDb webpage
- Number of user reviews
- Language
- Production country

- Rating
- Budget
- Year released
- IMDb Score
- Aspect ratio
- Likes on movie’s Facebook page
- Leading 3 actors and the number of likes on their Facebook pages

Modeling

We built our machine learning model in Weka, a data mining software platform with a large number of machine learning algorithms that can be applied to a dataset. In order to import our data into Weka, we needed to clean our data. We had to prune out characters that Weka couldn’t handle in our data (= “ ‘ * + - %). We also eliminated all examples that had no value for gross or budget, since these attributes were used to classify our data. 1,155 instances had no value for the film’s gross or budget.

We also removed attributes with information that would not be available prior to a film’s release. If our aim is to build a tool to predict a film’s success, it makes no sense to consider those attributes with information available only after a film has been released. We therefore did not consider the number of user reviews or number of user votes.

We also needed to classify our data. Since the aim of our project is to predict whether a film will be successful, we subtracted each film’s budget from its gross box office revenue. If the value were greater than zero, we considered the film a hit. Else, we considered the film a flop. Accordingly, we did not consider gross as an attribute when building our predictor.

Lastly, we needed a method for assessing the machine learning models we built. In this report, accuracy is the number of correctly classified instances divided by the total number of instances. Each model was assessed using 10-fold cross validation.

At this point, we built two models through Weka. Generating a pruned decision tree resulted in a tree in which a film was always designated a hit. Therefore, the accuracy of our tree was the same as the percentage of instances that are hits (55.76%). Using a 3-nearest neighbor approach, we achieved an accuracy of 62.62%. We attempted to build a random forest model of our data, but did not have the computational power. This accuracy was not encouraging, as it was not much better than simply regressing to the most common classification.

We then considered ways to optimize our data to make these algorithms more accurate. We started by deleting several attributes. The names of directors were too varied to be useful as nominal data and provided too little information to be useful as textual data. Since the most prolific director in our data set, Steven Spielberg, appeared only 26 times across our 4,160 examples, the name of the director had little impact on our model and only acted as noise. The same is true for the names of the leading actors, movie titles, and plot keywords. We deleted the attribute for a movie's IMDb website, since that link is unique to each film and has no bearing on a film's profitability.

These steps had little impact on the accuracy of the nearest neighbor approach, but increased the accuracy of the decision tree to 57.35%. The biggest improvement was that these changes greatly reduced the computational cost of producing these models, so we were able to generate a random forest model. This model had an accuracy of 62.33%.

Considering further ways to optimize our data, we targeted the nominal attributes. For instance, each movie had a series of genres listed under the genre attribute. Some films had six or seven genres listed. We had been treating each of these lists of genres as distinct values, so our model was considering 790 distinct genre values nominally. We modified this attribute so that each movie only had one associated genre (the first, since that was generally the most descriptive). Instead of considering 790 distinct genre values, our model now considered only 15.

We modified the language attribute as well. Instead of considering 41 possible values, we adjusted the parameter so the model considered whether the movie was English or foreign language. Similarly, we adjusted the country attribute to a boolean on whether the movie was produced domestically or internationally.

Conclusions

Following these changes, the accuracy of our decision tree model had increased to 65.93%. The accuracy of the 3-nearest neighbor model had increased to 63.08%. Most significantly, the accuracy of our random forest model had increased to 71.69%. We believe the random forest algorithm was the most useful since it classifies based on the results of a number of decision trees.

In analyzing our results, Weka also has a tool for analyzing which attributes are the most predictive in our random forest (based on the average impurity decrease). The five most predictive attributes are:

1. Number of critic reviews
2. Duration
3. Likes on director's Facebook page
4. Color
5. Likes on 3rd actor's Facebook page

Our team drew two conclusions from our research. First, the more publicized a film is, the more likely it is to be successful. Many of the attributes that prove highly predictable for a film are indirect measures of its popularity. A film with a large social media presence (likes on its Facebook page) or great critical analysis (number of critic reviews) is likely to fare better in the box office. Second, it is difficult to build a highly accurate classifier. While we are pleased with the progress we made over the course of this project, it is clear that there is much room for improvement with our classifier. An accuracy of 71.69% is encouraging, but there remain opportunities for improvement. Many of the our attributes are indirect measures of elements that may be important to a film's success. The number of likes on actors' Facebook pages is an indirect measure of the star power of a film. The IMDb score is an indirect measure of a film's quality. The elements that might be the most essential to a film- like a compelling story, a great cast, or a powerful message- are extremely difficult to collect data on and to use in a machine learning approach.