Predictive Analytics

11/15/2019

Introduction

The purpose of the project was to predict a movie rating score that would determine whether a movie is considered a good movie or not. The targeted user was online audio/video platforms such as Netflix, Amazon Prime Video, etc.

A logistic regression model was made after testing the 45000 movie rating found in the Full MovieLens Dataset consisting of movies released on or before July 2017.

The model includes movie popularity, runtime, released year, budget and production countries as influence factors. The effect of each variable to the movie rating score (good score) is explained in detail below along with a series of new findings.

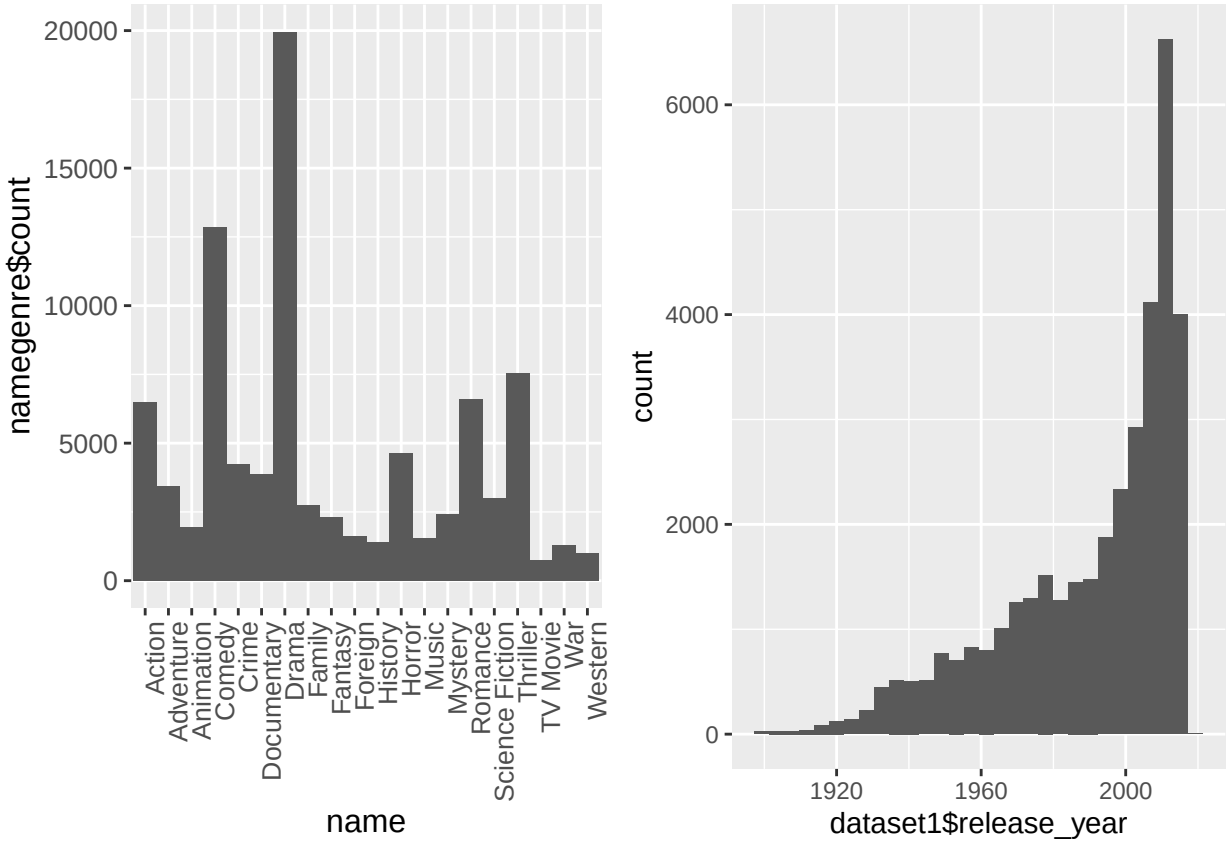
Data Interrogation / Analysis

After several iterations, we have observed some phenomena.

From the data, In terms of quantity, we can see that the entire movie market has produced a lot of movies in the past 100 years, and it has been developing with an upward trend, especially in the 21st century, which has accelerated the pace of development and reached the highest peak so far in 2014. At the peak, and then gradually began to slow down.

In terms of genre, it shows that the most common/popular top 5 genres out of the total 20 in the movie market are: Drama, Comedy, Thriller, Romance, Action. Their sum is even more than 50% of all movies.

According to those observations, we can roughly assume that with the increasing competition in the movie industry, the average return on capital is dropping down, which caused some capital is withdrawing from this market, but the return on capital that good movies can bring is still massive. This is also the meaning of this report: making business decisions from the prediction of whether it's a good movie or not.



Predictive Model Approach

In order for the team to properly select the appropriate features necessary to build the most optimal model, we tested out different models and finally decided using the model with the following variables include.

1. Popularity - A numeric quantity specifying the movie's popularity

   Popularity negatively influences the possibility of a movie getting a good score. The more widely and randomly the movie spread, the more likely that the movie will get a lower score. The range of popularity is 0 to 21.02. For example, when the popularity increases from 1 to 10, the possibility of a move getting a good score decrease by 66%. That is because, with the size of the population who watch the movie get larger, aesthetics and appetites become more dynamic and various.

   The more various the attitudes towards the movie is, the more likely the movie will get scores with a wide range from low to high. For example, a popular horror movie watched by an audience who prefers comedies is very likely getting a low score from that audience.

   One good solution for this problem is to send related notifications of movies to users with interests instead of sending notifications to all users.

2. Runtime - The running time of the movie in minutes

   It shows that the runtime of a movie also can negatively influence the possibility of a movie getting a good score. The range of runtime in the dataset is 0 to 209 minutes. When runtime increases from 1.5 hour (90 minutes) to 2 hours (120 minutes), the possibility of good score decrease by 5.71%; when runtime increases from 2 hours (120 minutes) to 2.5 hours (150 minutes), the possibility decrease by 5.30%; when runtime increases from 2.5 hours (150 minutes) to 3 hours (180 minutes), the possibility decrease by 4.92%.

   Although the marginal effect decreases to some extent, the possibility continues decreasing as runtime increases. It is common that users get bored and impatient for long movies. In that situation, users are very likely to give low scores. One solution is to provide users a function to stop and store watching history for the movie and get back to continue watching it later. Also, the abstract of the movie should be described accurately so that users can know whether the movie is the one they really want to watch.

3. Release year - The year on which it was released (get from the release_date)

   The year of release is positively related to the possibility of a movie getting good scores. In other words, movies that are released more recently are more likely to get good scores. That may be because movies nowadays do a better job of reflecting what people care about currently and contemporarily.

   The range of the year of release is 1900 to 2020. Compared to the movies released in 1990, movies released in 2020 are 7% more likely to get good scores. It suggests that contemporary movies are more worthwhile to purchased and promoted to customers.

4. Budget - The budget in which the movie was made

   We can see that the budget of movies will influence the average rating, therefore influencing our decision that whether they are good movies or not. However, our result shows that movies' budgets may have a negative effect on movies' ratings. If the filmmakers invest 1 dollar more in the movie, there will be 1.04 decrease in a good score. In other words, the more budget a movie has, the more likely that the probability of the good movie decreases.

   Maybe the "artistic cheapening" phenomenon accounts for this situation. The blockbuster may seem to use CGI technology to make up for a lack of plot and dialogue, leading to a decrease in the movies' quality.

   Therefore, because the average of movies' budget is 1047180 dollars, if a movie's budget is higher than approximately 1000,000 dollars, maybe there will be an increased likelihood that this movie is not a good movie.

5. Production_countries - The country in which it was produced

We can see that the production country has an influence on the probability of a good score of a movie. The result shows that if the movie's production country is not US, the probability to be a good movie would increase. For example, when we set popularity is 10, the runtime is 100, release year is 2000, budget is 40000, the probability of good score of movies with non-US production country would increase about 0.7% compared to the movie with US production country.

In this situation is because our investigate market is the US, so when they buy some foreign countries' movies, they would like to choose some movies with a high reputation and high score. Therefore, foreign movies in US marketing are actually some good movies and have been selected. So, compare to the local movies which are not selected, the probability of a good score of foreign movies is larger.

Although we successfully built the model, from the summary of GLM regression, we find that popularity, runtime and budget leave a negative influence on the good score, which is little inconsistent with common business meaning. However, the p-value for these variables is significant. AIC also is relatively smallest among other models. Residual plot shows little fitted discernible pattern, too. Therefore, from the class of endogeneity, we guess that our model may have some omit ted variables, which may have left some influence on dependent variables and good core. In order to solve this problem, we may need to set up an instrumental variable such as the number of actors, which is unlikely to directly influence the goodScore. But, more actors acted in movies may be more likely to increase the budget (correlated with budget, x) and use two-stage least squares to get an unbiased estimate of model.

Use Cases

After observing and predicting history movie data, we recommend that the users should choose the movies that have last 10% popularity, last 10% budget, last 10% runtime, top 10% release year and are produced in non-us countries considering the negative and positive effects on showing probability of good movie.

In other words, if the users see a non-us movie that has popularity lower than 0.078, budget equals to 0, runtime less than 62 and is later than 2014, there is relatively large 12% probability that this movie is a good movie and has the high average rating greater than 4.