Aashya Khanduja

24 May 2018

Logistic Regression for Predicting Movie's Success

Abstract

This project was carried out to correctly estimate the probability with which an unreleased movie will be successful in the given market. With further analysis and modification, this study can be used as early as in predicting if a movie's script will provide a favourable outcome for the production house.

Using predictive modelling, we identify patterns and anomalies in a data set comprising of historical information. This obtained order aids in foreseeing a value associated with new information and provides a probability attached to it. Its applications have proven to be invaluable in the field of social science, but we wish to extend it into further arenas. This paper presents steps involved in developing a Logistic Regression model based on various parameters affecting movie ticket sales.

Introduction

Despite logistic regression model being known for its predictive abilities for years, and being used for a number of predicting cases around the world,[3] it has rarely been used for improving the quality of entertainment we receive or to increase revenue for media industry. The emphasis of this project lies on increasing profits for the industry, and on giving quality entertainment to consumers.

Movie industry has been trying to make informed decisions about the genre and content of movie, by gathering information through surveys done by talking to general masses.[5] They are constantly facing challenges to provide us with movies having great content irrespective of its genre. An analysis of movies being hit based on genres is of particular interest to production houses and media firms. While some movies are extremely popular amongst people, others are unheard of by most. A large amount of effort is put into conceptualising, directing, and producing a single movie, whose success is determined by a plethora of factors that can and cannot be influenced by movie producers. The number of theatres available in a tier of cities, population residing in those tier cities, youth population, etc are critical factors that cannot be controlled by the industry. Diving deeper into the analysis we find the review of a movie to also be established from parameters like ticket amount, genre, series number, year of release, actors working in the film, voiceover actors, etc which are consequential to ticket sales and can be dictated by the movie maker.

An accurate predictive analysis model would be utilised by movie producers to use these features and maximise profits, thereby maintaining the customer base and company reputation. While appearing so on the surface, the impact of these movies being a failure or success is not restricted to the film industry, movie makers, and production houses, it extends far and wide into the world of consumers. Flop cinema that does not retrieve its cost and is not able to make its mark can also impact revenues of media companies, as a result leading to production of low budget films which may jeopardise the quality of movies being presented to the common man. Even a single movie doing badly has large impact on the image of that production house, and on the representation of that series.

Netflix has been a forerunner in deploying technology for providing movie and series choices based on consumer patterns .[1] Their model centred on data analytics, compiling relevant demographic and user preference information to predict appropriate selections, is much appreciated amongst citizens or netizens as it saves hours of search time. This data can be extended to the industry to narrow the target audience for movies, and even to lend a hand for creation of content that has the strength to uphold itself.

Along with a prediction model, a strong knowledge of the industry and an understanding of sentiment is also required. Netflix has often gone against data predictions where emotion and usability is in picture, such as when they made exit subscription process simpler by not burdening consumers with excess things.

Variables

Movie ticket prices are affected by a number of factors which vary with each segment, such as ticket prices being based on a theatre, and rating of actor being available for each actor working in that movie, therefore we take an average in such cases to neutralise the results. After identifying independent variables, an average of ticket prices, rating of actors working in that movie, youth population, year of release, release season, genre, whether it has received any awards, critic reviews, number of theatres where the movie was screened, if the movie is a sequel, we find the data with respect to a particular movie online. Data such as youth population for that year was not available for each particular year but an estimate could be found from the government population study database which is done once in half a decade or once in a decade.

Not taking movie name as a variable was done as ticket sales and the probability of that movie being profitable is only dependant on the name if the movie is a sequel, in which case we take the factor of sequel into consideration.

Table 1: Description of variables

Variable	**Definition**	**Characteristic**
Ticket Prices	An average of ticket prices for the movie	Number
Actors	Average rating of actors used	1 - A Rating, 2- B Rating, 3 - C Rating
Youth Population	Population of youth in targeted areas	Number

Year of Release	Year of Release	Number as an year
Release Season	If there is any holiday at the time of release	1 - Holiday, 2- No Holiday
Genre	Genre of the movie	1 - Animated, 2 - Drama, 3 - Romance, 4 - Action, 5 - Thriller
Awards	If the movie has won any awards	1 - Yes, 2 - Nomination only, 3 - No
Critic Reviews	Critic Reviews	1 - Great, 2- Good, 3 - Average, 4 - Bad, 5 - Very Bad
Theatres	Number of theatres	Number
Sequel	If the movie is a sequel	1 - Yes, 2 - No
Probablity	Probablity of the movie being a hit	1 - Yes, 2 - No

Logistic Regression

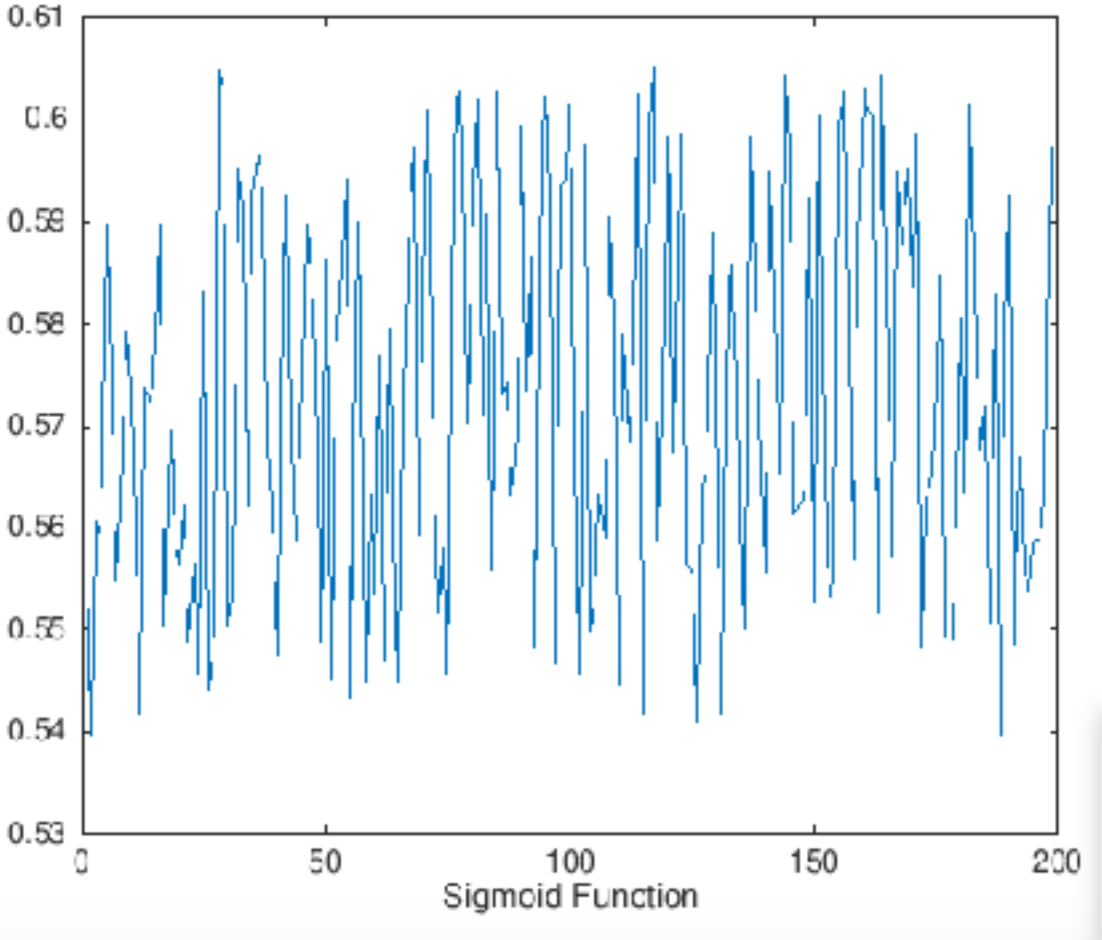
Logistic Regression (LR) is based on analysis of independent variables which then gives a dependant variable as an output as a dichotomous variable giving only 0 or 1.[2] The goal would be to predict the success of this movie if it goes on screen. These decisions are based on a number of tests done before choosing variables and factors. First and foremost we understand the problem at hand and see if it needs to be used for logistic regression or linear regression. A linear regression will give s continuous numbers as an output, which can be more than one. A linear regression would be used if we were trying to find the sales of a movie vs a logistic regression being used to check if probability of this movie being a hit. We perform model fitting test, by applying log likelihood, to check if we are using appropriate variables for a logistic regression. Classification is done to predict a value with some small discrete values as outputs.

The prediction through logistic regression is done with the help of sigmoid function which gives us the probability at a certain point. Sigmoid is found using simple mathematical calculations. As given below h gives the hypothesis value between 0 and 1 and g(z) computes sigmoid of z

```
h = sigmoid(X*theta);
```

g = 1 ./ (1 + (2.17).^-z);

We further see that for the dataset used, we get a sigmoid function giving probabilities as per the hypothesis value.



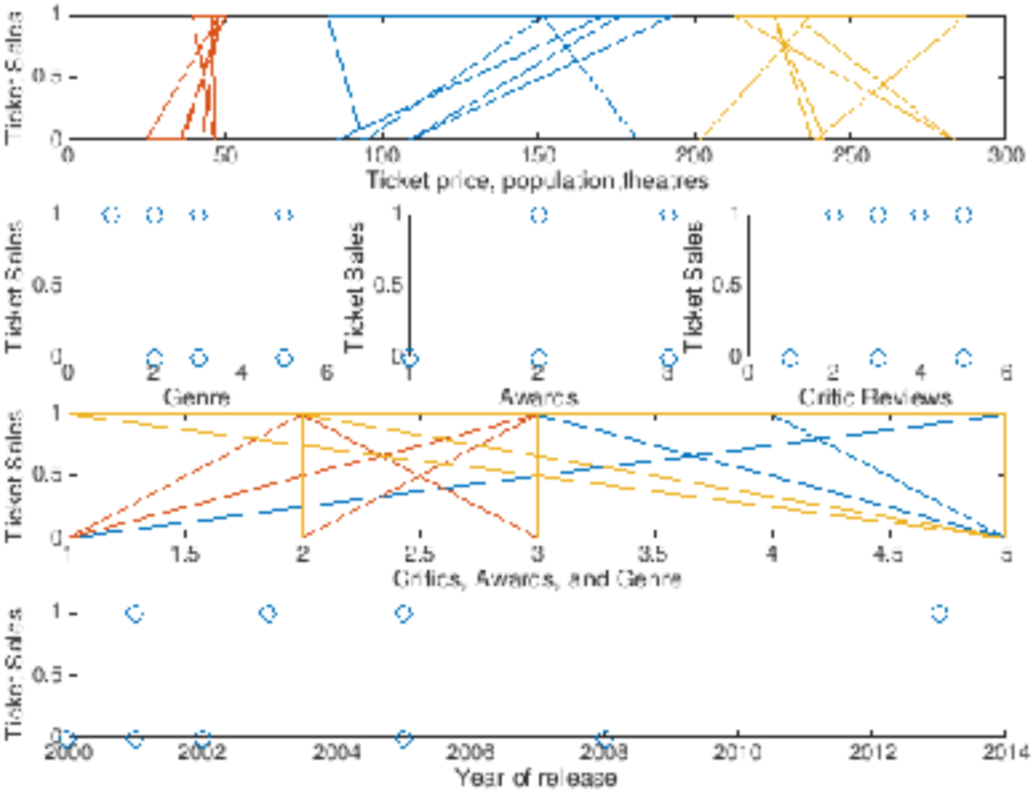
The objective of using logistic regression was to find the best model to describe the relationship between the dependent variable and independent variables. It generated coefficients to predict a logit transformation of the probability of success of movie released.[1]

Data Set Size

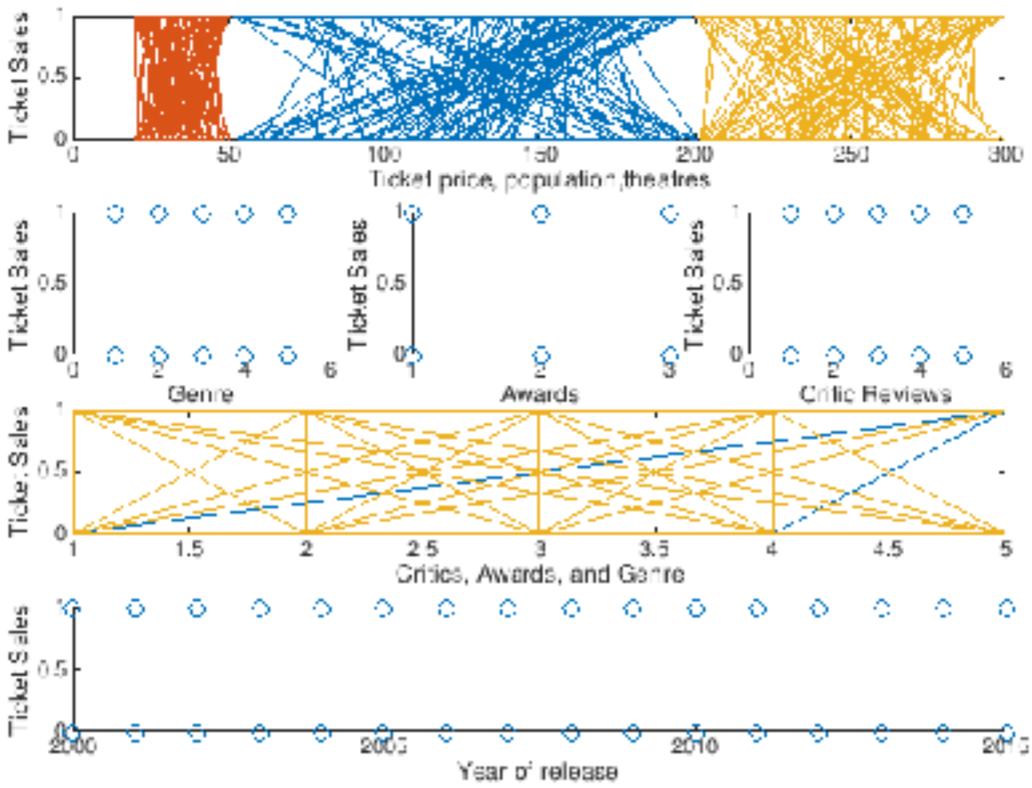
Taking a small (10 samples) and a large (200 samples) training dataset we have seen the correlation and the outcomes of logistic regression for the same kind of data.

We see the advantages of having a large training data set while making sure that we do not cause overfitting of data. Taking a small training set gives us very few variables to learn from and so the learning efficiency is low, hence giving us wrong outcomes. One of the basic requirements of machine learning is to have a large dataset that the algorithm can learn from and better itself. Giving a small training data set does not expose the algorithm to many test case and permutations and combinations, making the algorithm less effective.

While a smaller training data set gives us a clearer picture of the data available and where it lies on the graph, the relation between all aspects is not clearly visible. Using a smaller data set we can clearly determine the range of each of the values and see what probability a particular vale would give.



Taking a larger data set, we see correlations clearly but are visually unable to determine the probability associated with each value. As in the plot, we are able to see that when ticket prices(depicted in blue) are between 150 and 200 then the probability is more often 1 then when it is lower. We are also able to determine by looking at the graph that approximately 275-300 theatres (yellow) also give a higher probability than otherwise.



Another visual difference in taking a small or a large training set is that we see a clear distinction between critic reviews, awards obtained, and genre when we use a small training set, but when we shift to a larger training set, that distinction is not clear.

Through the predicted cost and gradient we observer that the initial cost and gradients are almost the same for both the training models (smaller and larger training set). At the test theta as well we see the cost to be same. But for theta found through fminunc (minimum unconstrained multivariate function), we observe a larger difference in the costs. While the cost at this theta for smaller training set is 0.159875, for larger training set we get 0.645031. We observed that thetas obtained herein also have a larger gap than gradients obtained at initial theta. A further and much crucial difference is seen when we predict the probability of success of the movie. Taking variables as (Ticket Price = 159, Actor rating = 2, Youth Population (in crores) = 34, Year of release = 2012, Holiday Release = 0 (No), Number of theatres = 289, Genre = 5 (Thriller), Awards = 3 (not won or nominated), Critics Reviews = 5, Sequel =   0 (no)), we predict a success rate probability of 0.00000 for smaller training data set and 0.627674 for larger training data set.

Vanishing Gradient vs Extremely low Gradient

During the process of implementation of logistic regression using gradient descent, we find gradient which is to converge and be minimised to give an optimal cost function. But sometimes, it converges much quicker than normal and we obtain very small values of gradient making the gradient almost vanish

While finding the gradient in this project, I obtained a value of -∞. Here, I thought that I am getting a vanishing gradient since it is an extremely small value, and tried resolving for the same.

A vanishing gradient will be an extremely small value close to 0 whereas -∞ is an extremely large negative value. The difference between the two becomes critical to be able to solve this and get better cost function. To solve a vanishing gradient error we wold need to shift from a sigmoid function to a Rectified Linear Unit as the root cause of the issue is the nature of sigmoid function, whereas to solve am extremely high value whether it is negative or positive, we need to normalise the input variables.

Conclusion

By following market trends, we see that the need for an algorithm such as this which can deduce the probability of a movie making profits is extremely necessary, moreover through studies that were performed we see that this prediction method is successful in the industry. Movie preferences are highly governed by geography and so no one movie can be liked by all. We find that some movies are liked by the general masses all over the globe barring their sexuality, religion or geography.

With further information, experiments have shown that this method can be sed to predict if the script of the movie is good and what are the movie genre preferences based on region, for example, action movies with superheroes are liked in US. With an even deeper dive into this model with more data, we can improve the quality of movies right from the script by analysing which part of the script is not liked by people. As the paper shows, features that have been chosen in machine learning had an impact on the prediction accuracy; however, some features are redundant predictors and can and need to be removed by algorithms.

The experiment also compared the prediction results between small and large data sets and showed that accuracy rate is positively correlated to a larger data set, but only to a point where the data is not overfitted. Excessive predictors can lead to overfitting and negatively influencing the prediction accuracy.

In my future research on detecting accuracy through screenplay and improving it, the attributes used will need to be changed to variables relating the screenplay. A large number of features related to screenplay will need to be detected and reviewed for being incorporated to build more accurate recommendation model.

References

1. Movie Genre Preference Prediction Using Machine Learning for Customer-Based Information — Haifeng Wang & Haili Zhang

2. A Logistic Regression Based Approach for Software Test Management — Yue Zhou & Jinyao Yan

3. BREAST CANCER ANALYSIS USING LOGISTIC REGRESSION H. — Yusuff, N. Mohamad, U.K. Ngah & A.S. Yahaya

4. An Introduction to Logistic Regression Analysis and Reporting — CHAO-YING JOANNE PENG, KUK LIDA LEE, GARY M. INGERSOLL

5. A LOGISTIC REGRESSION MODEL TO PREDICT FRESHMEN ENROLLMENTs — Vijayalakshmi Sampath, Andrew Flagel, Carolina Figueroa