Project Compiled by :

3-14-11-0414

3-14-11-0417

3-14-11-0419

3-14-11-0422

3-14-11-0432

3-14-11-0434

3-14-11-0451

PREDICTION OF OPENING WEEKEND BOX OFFICE PERFORMANCE

INTRODUCTION:

The term **'Box office'** is used in the context of the film industry. It refers to the amount of business a particular movie makes .The movie industry is a business with a high profile, and a highly variable revenue stream.

In 1998, moviegoers spent $6.88 billion at the U.S. box office alone.

A single movie can be the difference between millions of dollars of profits or losses for a studio in a given year. It's not surprising, therefore, that movie studios are intensely interested in predicting revenues from movies; the popular nature of the product results in great interest in gross revenues from the general public as well.

The statistical analysis of the box office revenue is also highly important for the creative industries and also a source of interest for the fans.

As we know, producing movie is one of the riskiest businesses, the movie industry being highly dynamic. The involvement of various factors makes it more ambiguous. Here our objective is to set up a model for predicting the box office success.

OBJECTIVE:

We know that the opening weekend box office performance of a particular movie largely determines its total revenue. Hence for simplicity, we try to predict only the opening weekend performance of a movie.

Now we know that different type of films has different classes of audiences. Thus taking movies without considering their genre (type) will not be meaningful.

Hence our objective is to find different models for predicting the box office success for different types of movies.

DATA:

Source of data:

For the analysis, we have taken into account Hollywood movies only that have released in USA in the time period 2012-13. We have collected our data from the following websites:

www.boxofficemojo.com

www.imdb.com

www.metacritic.com

Description of variables used:

Response:

Our response is the opening weekend collection of a movie which is a continuous variable.

For the prediction purpose, we have taken into account the following covariates:

1.Production budget:

Production budget of a film has been treated as a continuous variable.

2. No of screens:

It refers to the number of theatres in which the movie is released in its opening weekend

(In our data, we have taken the USA release only).

3.Critic ratings:

As a measure for critic ratings, we have used 'Metacritic score' given by the website www.metacritic.com. The website actually aggregates reviews of the films. For each film, a numerical score from each review is obtained and it is converted into percentage. Then a weighted average of the scores is done, weights being decided based on the critic's fame or stature. The score is given out of 100. It is a continuous variable.

4.Star power:

The concept ‘Star power’ captures the extent to which an artist’s involvement with an entertainment product contributes to the success of that product. We are to develop a measure for calculating the star power of the three leading actors (as dictated by ***www.imdb.com***) of a movie. We have assigned weights to the actors on the basis of the awards they have got or have got nominated for. The ordinal categories and the corresponding scores are shown below:

Awards	***Scores***
Oscar (won)	***8***
Oscar (nominated)	***7***
Golden globe (won)	***6***
Golden globe (nominated)	***5***
Prime time Emmy (won)	***4***
Prime time Emmy (nominated)	***3***
Bafta (won)	***2***
Bafta (nominated)	***1***

If an actor does not fall in any of the above categories, his/her star power has been taken as zero.

The star power of a movie will be equal to the sum of the star powers of the three leading actors of that movie.

For example, if in a film, we have three leading actors, one nominated for an Oscar and the other two possessing two Golden globes each, then the star power of that movie will be

(7+(2*6)+(2*6))=31.

5.Release timing:

It is an indicator variable in our analysis. It takes the value 1 if the movie is released in a festive weekend and zero otherwise. Festive seasons we have considered here are:

(a)New year's day

(b)Martin Luther king day.

(c)Memorial day

(d)Halloween.

(e)Thanksgiving day.

(f)Christmas.

6.Sequel:

It is also an indicator variable denoting whether a movie has a prequel or not, taking values 1 and 0 respectively.

PREPARATION OF DATA FOR ANALYSIS:

For any kind of analysis, we must check whether our response variable follows a normal distribution or not. We first perform the Shapiro-Wilk's normality test.

SHAPIRO WILK NORMALITY TEST:

Let $x_1,x_2,...,x_n$ be random sample of size n...

Here we are testing

H_0: $x_1,x_2,...,x_n$ came from a normal population

H_1:not H_0

The test statistic is given by:

W=$\frac{\sum_{i=1}^{n}\{a_i X_{(i)}\}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$

$X_{(i)}$=i^th^ order statistic ; i=1,2,..,n

$\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$

$(a_1,a_2,..a_n)$=$\frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m^T}}$

$m= (m_1,m_2,..,m_n)^T$

$m_1,m_2,\ldots,m_n$ are the expected values of the order statistics of an i.i.d. random variables sampled from standard normal distribution. V is the covariance of those order statistics.

We reject H_0 if W is below a threshold value.

p-value:

If p-value of this test is less than the chosen α level we reject null hypothesis... Here we choose α to be 0.05

APPLICATION :

At first we applied Shapiro- Wilk's test on our response y (opening weekend box office collection). The p-value obtained in this test is = **1.136e-13(<0.05)**

From the p-value, we see that the null hypothesis is rejected. Hence we should use some kind of transformation to normalize our response variable. At first we go for logarithmic transformation and observe that the p value increases slightly (**p–value : 6.248e-09**) but clearly less than 0.05.

Normality of the response is still not achieved.

Then we opt for square root transformation and we see that the p-value increases further (**p-value = 5.707e-05**) .Then we go on decreasing the power of the response. Finally, taking fourth root transformation, we get a p value which is equal to **0.6187(>0.05).**

We further decrease the power of the response variable and take eighth root.

We perform the normality test and again and the p value comes out to be **0.001996** which is less than 0.05. Hence we have got the maximum p value in case of the fourth root transformation**. So, we decide to take $\sqrt[4]{y}$ as our required transformation.**

The histogram also gives a good normality shape. Since our response takes positive values only, $\sqrt[4]{y}$ is a monotone transformation. Taking this transformation we proceed for further analysis.

CHOICE OF MOVIES :

We selected the time period 2012-13 for our analysis and we collect our data from the movies of this time interval.

By simple random sampling method we selected 120 movies from all the movies of 2012-2013.

Unfortunately data for three of the movies were missing, so finally we collected data from these 117 movies to do our analysis.

Before doing so we at first divided the movies on the basis of genre since we know that different type of films has different classes of audience and hence the analysis of all the films taking together will not be meaningful.

So we need to classify the selected films into a number of mutually exclusive and exhaustive genres. So we divide the movies into five major genres namely

(A) action

(B) animation

(C) comedy

(D) drama

(E) thriller

(combining the films of similar and overlapping genres into one genre

i.e. ,

Action (combining action and adventure movies)

Comedy (combining romance and comedy)

Drama (combining history, politics and drama)

Thriller (combining horror , thriller, mystery movies)

etc.)

Classification of the movies according to genre has been done with the help of **www.imdb.com.**