IIM AMRITSAR

A Project Report on

FACTORS DETERMINING THE IMDB RATING OF A MOVIE

Quantitative Methods

(GROUP - 8)

Submitted by:

Himanshu Goyal (MBA/06/078)

Varun Nair (MBA/06/088)

Prachi Parekh (MBA/06/091)

Yashasvini Mathur (MBA/06/113)

Submitted to:

Dr. Deepa Mishra

Abstract

Background: In this rapid era of digital world, there are so many ways by which everyone can entertain themselves and one of the ways is by virtual entertainment but even in this segment there are many options which can confuse anyone but still everybody knows one way to know which movie/series/game to watch/play. All of this is made easier by IMDB rating which everybody looks on before choosing anything. So that means how much important IMDB (Internet Movie Database) is for maximum of us. Even though IMDB takes online vote from its online users still we all believe on IMDB rating from 29 years.

Objectives: The primary objective of this study is to examine the effect of various factors on the IMDB rating of a movie by observing the correlation and using the factors having higher correlation for predicting the IMDB rating of a movie using a multiple regression equation. The secondary objective of this study is to analyze the ratings of each genre.

Subject and Methods: A Secondary data, "IMDB 5000 Movie dataset" is taken from Kaggle and analyzed based using the numerical attributes of the dataset due to the constraints of the software used i.e. MS Excel. Data was cleaned manually and missing values were eliminated from the data. A sample was selected from the data using Convenience Sampling – Data for 10 years (2006-2016) was used for analysis which included records of 1674 movies. The following tests were performed on the collected data: General Max-Min Analysis, Descriptive Statistics, Correlation Analysis and Regression analysis.

Result: - The main finding of the study is that, after performing Correlation Analysis the IMDB rating of a movie depends on the following 5 factors –

- num_critic_for_reviews
- duration
- num_voted_users
- num_user_for_reviews
- movie_facebook_likes

After applying multiple regression on the numeric data, the regression equation obtained is:-

$$Y = 4.59 + (0.0025)x_1 + (0.011)x_2 + (3.25\times10^{-6})x_3 + (-0.0012)x_4 + (-1.5\times10^{-6})x_5$$

Where

Y = Predicted IMDB Rating of movie

X_1 = num_critic_for_reviews

X_2 = duration

X_3 = num_voted_users

X_4 = movie_facebook_likes

The secondary findings of the study are: -

- Highest Rated Genre – Sci - Fi
- Lowest Rated Genre – Documentary
- Highest Rated Movie – The Lake HouseÂ
- Lowest Rated Movie – The Great RaidÂ

Literature Review

IMDB (Internet Movie Database) is an online database of information related to films, television programmes, video games, information about each and everything related to any segment that IMDB covers.

Currently, IMDB has approximately 6.5 million titles and 10.4 million personalities in database and 83 million registered users. Most data are contributed by volunteer contribution from registered users, without registration IMDB doesn't allow to share information.

IMDB allows users to rate on the scale of 10. IMDB indicates that the submitted ratings are filtered, and weighted in various ways and weighted mean rating is calculated and that becomes the final rating for the movie.

IMDB maintains a list of Top 250 movies in which movies are included on the basis of votes received from the regular voters which is kept as a discretion by IMDB who they consider as regular voters. In additions to other weightings The Top 250 films are also based upon credibility formula. Current formula is not disclosed but earlier the formula used by IMDB was

$$W= (R\times v + C\times m) / (v + m)$$

Where,

- W= weighted rating
- R= Average for the movie as a number from 1 to 10 (mean)
- v= Number of the votes for the movie
- C= mean vote across whole report
- m= minimum votes required to be listed in Top 250

All the contributors of the database technically retain copyright on their contributions, but the compilation of the content becomes the exclusive property of IMDB with the full right to copy, modify and sublicense it, and are verified before posting.

Methods Used

- **Description of Data:** The dataset used is the "IMDB 5000 Movie dataset" taken from the open source repository Kaggle. The data before cleaning had 5675 rows and 28 attributes. The data was cleaned manually and incorrect values were eliminated. The attributes present in the data were :-

  Categorical Attributes:-

1. color – Black and White or Color Movie
2. director_name – Name of director of Movie
3. actor_1_name – Name of first actor in the movie
4. actor_2_name – Name of second actor in the movie
5. actor_3_name – Name of third actor in the movie
6. movie_title – Name of the movie
7. Genre – Movie belongs to which genre
8. plot_keywords – Keywords using which movie could be searched on IMDB
9. movie_imdb_link – Link of IMDB website where movie rating could be found
10. language – Language in which the movie is made
11. country – Country where the movie was made
12. content_rating – Content Rating of the movie

13. title_year – Year of release of the movie

Some movies belonged to multiple genres, so for the sake of simplicity, movies that belonged to multiple genres were assigned the genre that was written first in the "genre" attribute of their details.

Numerical Attributes: -

1. num_critic_for_reviews – Number of critics who have reviewed the movie
2. duration – Running length of the movie
3. director_facebook_likes – Number of likes on the Facebook page of the director of the movie
4. cast_total _facebook_likes – Total likes on the Facebook page of all actors of a movie
5. actor_1_facebook_likes – Number of likes on the Facebook page of first actor of the movie
6. actor_2_facebook_likes - Number of likes on the Facebook page of second actor of the movie
7. actor_3_facebook_likes - Number of likes on the Facebook page of second actor of the movie
8. movie_facebook_likes – Number of liked on the Facebook page of the movie
9. gross – Gross profit at the Box Office earned by the movie
10. facenumber_in_poster – Number of actors appearing in the poster
11. num_user_for_reviews – Number of users that have reviewed the movie
12. num_voted_users – Number of users who have voted for the movie
13. budget – Budget of the movie
14. aspect_ratio – Aspect ratio of the movie
15. imdb_score – IMDB rating of the movie

Due to constraints of the software used for analysis i.e. MS Excel, only numerical data was used for regression and correlation analysis.

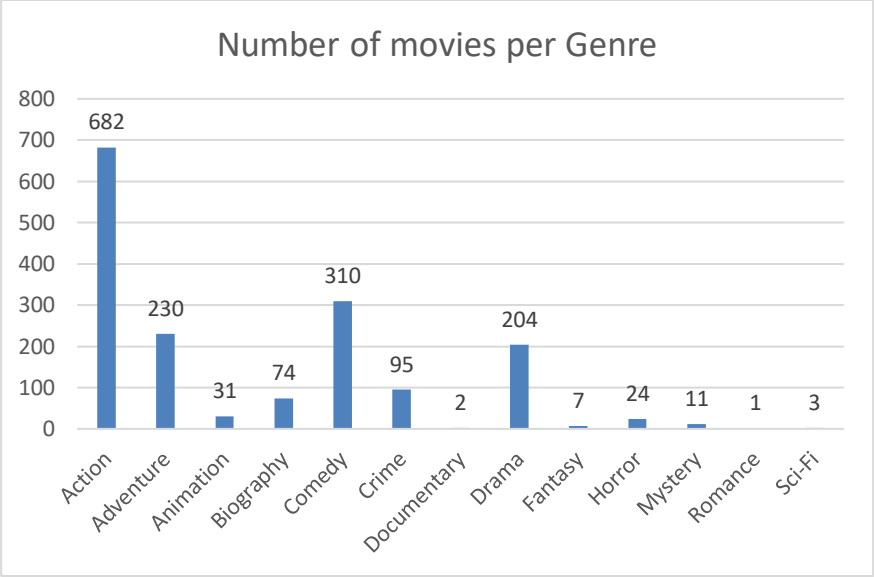
A sample was selected from the data using Convenience Sampling – Data for 11 years (2006-2016) was used for analysis. This sample include records of 1674 movies.

- **General Max-Min Analysis** – This included dividing the data category wise based on genres and years of release and examining the highest and lowest IMDB ratings per genre, genres having highest and lowest average rating and movies having highest and lowest rating
- **Descriptive Statistics -** It is the brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. It is broken down into measures of central tendency (mean, median, mode) and measure of variability (standard deviation, variance, maximum and minimum variables)
- **Correlation**- It is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.
- **Regression Analysis**- Regression analysis is a powerful statistical method that allows us to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

- **Software Used-** The analysis has been performed in Microsoft Excel 2019.

Results and Analysis

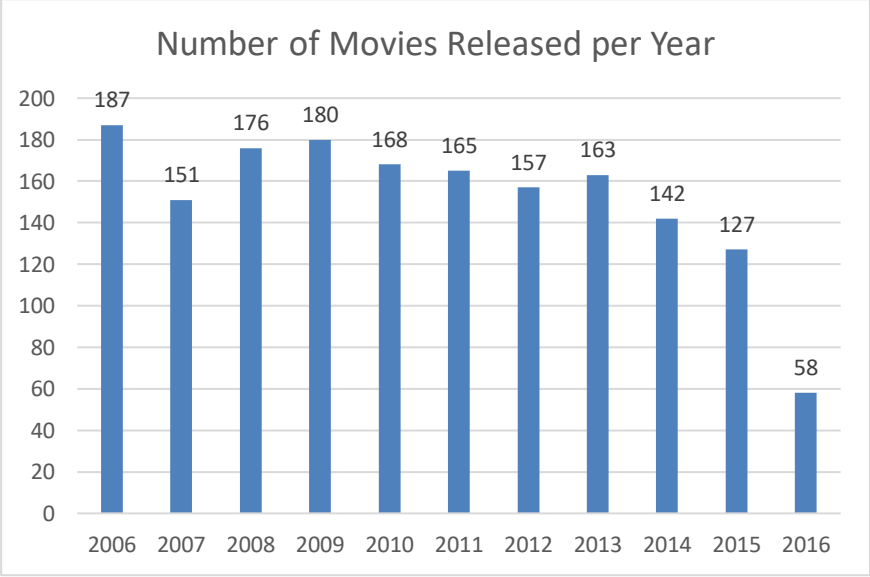
Graph 1



Graph 1 shows the distribution number of movies released per genre through the years 2006-2016.

It can be observed that Action movies were released the most and

Graph 2



Graph 2 shows the distribution of movies released per year through the years 2006-2016.

It can be observed that the highest number of movies released during 2006 and lowest during 2016.

Table 1

Genre	Average of imdb_score
Action	6.447507331
Adventure	6.447826087
Animation	6.683870968
Biography	6.32972973
Comedy	6.332258065
Crime	6.541052632
Documentary	5.7
Drama	6.471568627
Fantasy	6.228571429
Horror	6.45
Mystery	6.827272727
Romance	6.4
Sci-Fi	7.633333333

Table 1 shows the average IMDB rating of movies of each genre. From this we deduced that movies belonging to the Documentary genre have the lowest average IMDB rating of 5.7 and movies belonging to Sci-Fi have the highest average IMDB rating of 7.63.

Taking reference from Graph 1, we can observe that only 2 movies of Documentary genre are considered for the average and 3 movies of Sci-Fi genre are considered for the average, we can deduce that the average is highly affected by the number of observations used. There are cases in other genres where the number of movies were more and the average rating is a more reliable factor for consideration for evaluating the likeness of a genre by

Table 2

Genre	Max of imdb_score	Min of imdb_score
Action	8.6	1.6
Adventure	8.5	2.7
Animation	8.8	4.4
Biography	8.5	2.8
Comedy	8.5	1.9
Crime	8.3	3.6
Documentary	6	5.4
Drama	9	2.3
Fantasy	7.2	5.1
Horror	7.9	4.4
Mystery	8.4	5.4
Romance	6.4	6.4
Sci-Fi	8.5	7

Table 2 shows the IMDB score of highest and lowest rated movie of each genre.

The Lake HouseÂ is the highest rated movie of IMDB score 9 and belongs to the Drama genre.

The Great RaidÂ is the lowest rated movie of IMDB score 1.6 and belongs to Action genre.

We can observe that the highest and lowest movie do not belong to either of the genres having highest and lowest average IMDB score. This explains that averages are not useful for analyzing the

Table 3

imdb_score	
Mean	6.43644
Standard Error	0.024731
Median	6.5
Mode	6.7
Standard Deviation	1.011859
Sample Variance	1.02386
Kurtosis	1.40656
Skewness	-0.80777
Range	7.4
Minimum	1.6
Maximum	9
Sum	10774.6
Count	1674

Table 3 shows the Descriptive Statistical analysis of the IMDB scores of movies through the years 2006-2016. The average IMDB rating of movies in these years is 6.43 with median rating of 6.5 with a standard deviation of 1.012 and the variance is 1.02. The sample is left skewed with a skewness of -0.8 and has a higher peak than normal distribution curve with a kurtosis of 1.4. The lowest IMDB rating of a movie is 1.6 and the highest is 9.

Table 4

	num_critic_for_reviews	*duration*	*director_facebook_likes*	*actor_3_facebook_likes*	*actor_1_facebook_likes*	*gross*	*num_voted_users*	*cast_total_facebook_likes*	*facenumber_in_poster*	*num_user_for_reviews*	*budget*	*actor_2_facebook_likes*	*movie_facebook_likes*	*aspect_ratio*	*imdb_score*
num_critic_for_reviews	1														
duration	0.41365501	1													
director_facebook_likes	0.27352722	0.25047286	1												
actor_3_facebook_likes	0.24820183	0.19091218	0.17074217	1											
actor_1_facebook_likes	0.18761798	0.13623812	0.113139173	0.273231074	1										
gross	0.53584764	0.31773263	0.164861462	0.355740703	0.180969793	1									
num_voted_users	0.74206848	0.42996214	0.367411431	0.347236248	0.215674672	0.68522694	1								
cast_total_facebook_likes	0.26219054	0.18816028	0.159121174	0.563140924	0.916004894	0.303105061	0.315636007	1							
facenumber_in_poster	-0.08349955	0.04919559	-0.06643536	0.120471962	0.040182888	-0.04556673	-0.054892155	0.072889236	1						
num_user_for_reviews	0.6986084	0.44863737	0.30027633	0.269099425	0.160297572	0.625179122	0.814957272	0.24283386	-0.10547309	1					
budget	0.10057274	0.07201738	0.023606283	0.04055777	0.022671787	0.10613492	0.07792032	0.036031822	-0.01933346	0.085596578	1				
actor_2_facebook_likes	0.24897039	0.15408341	0.151093125	0.552420142	0.495119169	0.283962476	0.284292963	0.763384615	0.071791323	0.219929184	0.034919	1			
movie_facebook_l	0.72274508	0.37964338	0.242773684	0.272819832	0.161277389	0.446081515	0.619828119	0.241618403	-0.01340241	0.522200482	0.048707	0.227554206	1		
aspect_ratio	0.05913806	0.12617527	0.039165471	0.016411604	0.037551938	0.022971546	0.057511014	0.042129448	-0.00275587	0.061707663	-0.00339	0.042742604	0.06003036	1	
imdb_score	0.47201522	0.36267333	0.194027035	0.077638234	0.078455482	0.226348787	0.473521162	0.096111679	-0.10739672	0.291250349	0.035748	0.095652268	0.37256426	0.06014	1

Table 4 is the correlation matrix of the sample. The correlation coefficient of each attribute w.r.t IMDB score is calculated. Attributes having positive correlation coefficients increase the IMDB rating of a movie and the top 5 are taken into

consideration for regression analysis for deducing the equation for prediction of IMDB score. They are –

- num_critic_for_reviews
- duration
- num_voted_users
- num_user_for_reviews
- movie_facebook_likes

Since the correlation coefficient of num_voted_users is the highest, it influences the IMDB score the most.

Here we also observe that none of the attributes have a very high positive or negative correlation coefficient i.e. greater that 0.8 or less than -0.8. This explains that the IMDB score not only depends on the numerical attributes but is also affected by the categorical attributes. Due to the constraints of the platform used, categorical data could not be used for consideration during Correlation Analysis.

Table 5

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.5808065
R Square	0.3373362
Adjusted R Square	0.3353498
Standard Error	0.8249294
Observations	1674

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	577.829042	115.5658	169.8227	3.2453E-146
Residual	1668	1135.088138	0.680508		
Total	1673	1712.91718			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.5938682	0.129792963	35.39382	3.7E-205	4.33929398	4.848442	4.33929398	4.848442497
num_critic_for_reviews	0.0024829	0.000259245	9.577589	3.44E-21	0.001974464	0.002991	0.001974464	0.002991424
duration	0.0114815	0.001219117	9.41786	1.48E-20	0.009090311	0.013873	0.009090311	0.013872632
num_voted_users	3.522E-06	2.6565E-07	13.25644	3.3E-38	3.00053E-06	4.04E-06	3.00053E-06	4.04261E-06
num_user_for_reviews	-0.001185	9.78262E-05	-12.1091	2.07E-32	-0.00137646	-0.000993	-0.001376458	-0.00099271
movie_facebook_likes	-1.54E-06	1.03579E-06	-1.49127	0.13608	-3.5762E-06	4.87E-07	-3.57622E-06	4.86943E-07

Table 5 is the Regression Analysis of the top 5 selected attributes. The regression equation obtained is –

$$Y = 4.59 + (0.0025)x_1 + (0.011)x_2 + (3.25\times10^{-6})x_3 + (-0.0012)x_4 + (-1.5\times10^{-6})x_5$$

The Adjusted R Square is 33.53% which explains that approximately 33.53% of the observed variation can be explained by the model's inputs. This is rather lower proportion and can be explained by the fact that not just numerical attributes affect the prediction of IMDB rating, but categorical variables also do so. If we use a platform that can effectively measure the correlation coefficient of categorical variables also, it would be useful for making a more efficient model for prediction.

Conclusion

- Documentary genre has the lowest average IMDB rating of 5.7 and Sci-Fi has the highest of 7.63
- Movies of action genre have the highest number of releases and also an average rating of 6.4 which means that people are fairly fond of action movies.
- The IMDB score is not just affected by the numerical attributes of the given data, but is also affected by the categorical attributes, which if considered could have resulted in a more efficient model of prediction.

References

[1] IMDB 5000 Movie Dataset (2017). Retrieved from :
https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

[2] IMDb (2018) Retrieved from :
https://en.wikipedia.org/wiki/IMDb