International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952102

Multi-genre Movie Data Analysis Using Pearson's Correlation

Shalini Mounika Maturu*, Tamma Eekshitha, Talla Ramya, Tulluru Thanmai Durga
Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

ABSTRACT

The success of a movie plays an important role because it usually involves huge amounts of investments. Thus it becomes important to know beforehand whether the movie will be successful or not. The aim of our work is to prove that various attributes or factors related to a movie could prove useful in predicting the success or failure of a movie. Since one single attribute is not sufficient to predict the success of a movie, we've used multiple attributes and the comparisons between various attributes and their correlation for the success prediction. Therefore, in this paper we are considering a statistical technique called Pearson's correlation coefficient in finding out which factors are highly correlated with a movie's ratings.

Keywords: Movies, Success, Attributes, Pearson's Correlation.

I. INTRODUCTION

Huge investments go into the making of a movie therefore making it important for the movie to be successful. Knowing which movies are likely to succeed and which are likely to fail before the release could benefit the production houses greatly as it will enable them to focus their advertising campaigns as well which again require huge amounts of money, accordingly. So, we can say that advertising campaigns contribute heavily to the total budget of the movies. Sometimes the investment results in heavy losses to the producers. If it was somehow possible to know beforehand the likelihood of success of the movies, the production houses could be saved from such losses and also gain maximum profit.

That is the reason why forecasting box-office revenues of a particular movie has intrigued many scholars and industry leaders as a difficult and challenging problem. Despite the difficulty associated with the unpredictable nature of the problem domain, most analysts have tried to predict the total box-office receipt of motion pictures after a movie's initial theatrical release. Several researchers have attempted to develop models for forecasting the financial success of motion pictures, primarily using statistics-based forecasting approaches. Because, statistical approaches are quite useful for predictive purposes in domains like stock markets where again, investments made could result in profit or loss. Hence, in this paper we propose statistical model called Pearson's correlation coefficient. The Pearson's correlation analysis is used to find the correlation between two attributes. In other words, to find whether one factor impacts the other factor or not (positively or negatively). In our work we tried to prove that various attributes or factors related to a movie could prove useful in predicting the success or failure of a movie. Since one single attribute is not sufficient to predict the success of a movie, we considered multiple attributes and the comparisons between various attributes and their correlation for the success prediction. We used the Pearson's correlation technique to find out which factors are highly correlated with a movie's ratings.

CSEIT1952102 | Received : 15 March 2019 | Accepted : 03 April 2019 | March-April -2019 [5 (2) : 881-887]

For our research work, we considered the IMDB dataset. Internet Movie Database (IMDb) is a free, user-maintained, online resource of production details for over 390,000 movies, television series and video games, which contains information such as title, genre, box-office taking, cast credits and user's ratings. Hence making the IMDB is an excellent resource to find detailed information about almost any film ever made. It contains a vast amount of data, which undoubtedly contains much valuable information about general trends in films. The IMDB dataset is very useful for performing data analysis which helps us to uncover information which will both confirm or disprove common assumptions about movies.

II. RELATED WORK

In 2006, Ramesh Sharda and Dursun Delen performed analysis of prediction box-office movies success of motion pictures, forecasting performance of a movie at the box-office before theatrical release. To categorize movie whether flop or blockbuster, classification approach is used. These all information is from the IMDB data and also news data for the better performance. In 2015, Parag Ahivele and Omkar Acharya performed success prediction of films at box office using machine learning. In this Linear regression techniques are used to find the relation between the variables to predict the success. In 2004, Saraee, MH, White, S and Eccleston J performed ‘A data mining approach to analysis and prediction of movie ratings’ purely data mining techniques are used for the prediction of the success. In 2018, K Meenakshi, G Maragatham, Neha Agarwal and Ishitha Ghosh are performed ‘A Data Mining Technique for Analyzing and Predicting the success of Movie’(hollywood), here they used K-means clustering approach. In 2014, Nithin VR, Pranav M, Sarath Babu PB, Lijiya A performed analysis on the Predicting Movie Success Based on IMDB Data, where directly by taking the IMDb data is not good because the format problem,so step by step data processing is done and regression models(Linear regression model, Logistic regression model, Support vector Machine Regression Model) are used for the success prediction.

III. PROPOSED MODEL

Algorithm for movie success prediction:

a) Data is preprocessed, cleaned, integrated and transformed to simulation data.
b) Perform correlation analysis between genres and ratings.
c) Perform correlation analysis between particular genres.
d) Perform correlation analysis between actors and the ratings of their movies.
e) Now predict the success of the particular movie from the correlations.

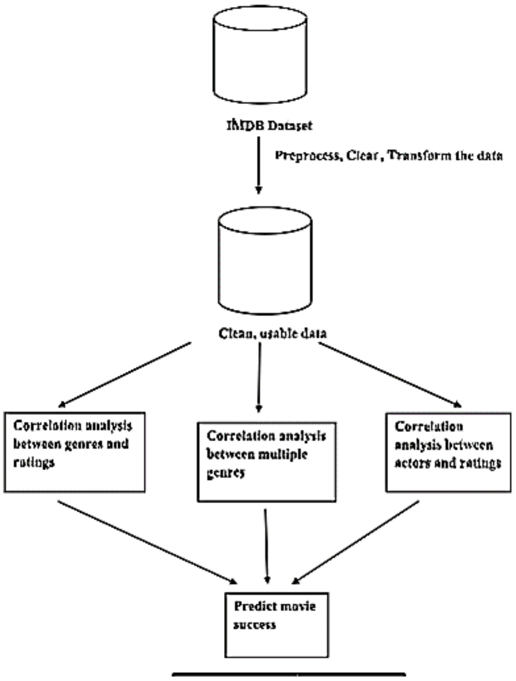

IMDB Dataset
Preprocess, Clear, Transform the data
Clean, usable data
Correlation analysis between genres and ratings
Correlation analysis between multiple genres
Correlation analysis between actors and ratings
Predict movie success

Fig 1: Data Cleaning Procedure

IV. DATA COLLECTION AND PRE-PROCESSING

A. Dataset Collection:

The initial dataset to be used is collected from IMDB. It consists of movies that were released from 2006 to 2016. Out of those movies, we selected movies which are in English since we considered the Hollywood dataset primarily. We got data regarding 1050 films.

B. Data Pre-processing:

We removed movies which don't have any information about box office details. The data we obtained are highly susceptible to noisy, missing and inconsistent data due to the huge size and their likely origin from multiple, heterogeneous sources. The main problem with datasets was some missing fields. Since, only few fields were missing. We removed those rows which had any empty fields.

C. Data Transformation

Since the dataset had many features or attributes, we transformed the required columns into a new file for ease of performing the analysis on two features at a time.

V. IMPLEMENTATION

In our work, we used a statistical model called Pearson's correlation analysis which is used for finding the correlation between two variables. Correlation 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. Now, the aim of our work is to prove that various attributes or factors related to a movie could prove useful in predicting the success or failure of a movie. Since one single attribute is not sufficient to predict the success of a movie, we have considered multiple attributes and performed correlation analysis on them and analyzed the correlation coefficient.

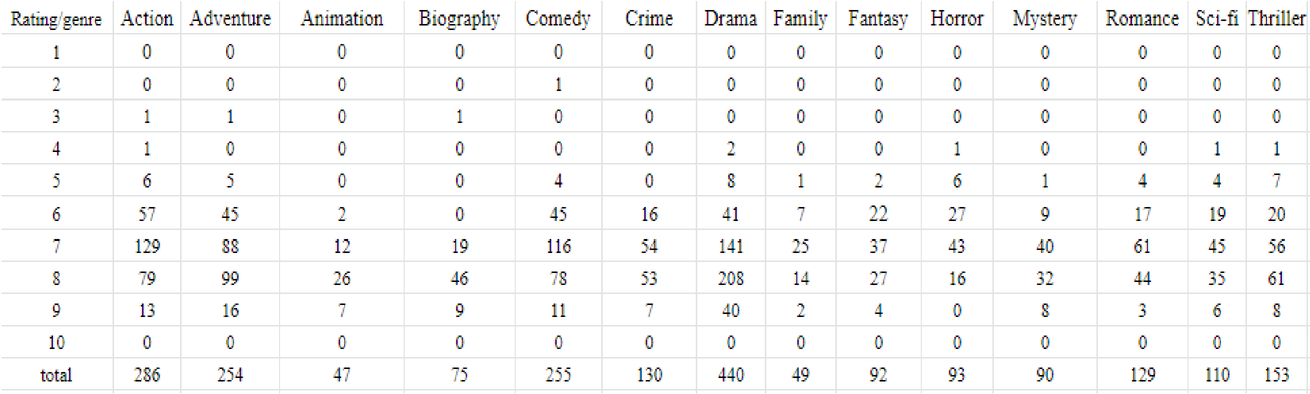
a)Tools used:

We used the inbuilt method "chi2_contingency(table)" from scipy package in python for obtaining the expected frequencies table. We performed the correlation analysis in python. We used MS-Excel for creating the graphs.

b) ***Genres vs ratings:***

After the data is preprocessed, cleaned and transformed, we considered only the genres and their rating values to find out the correlation between them.

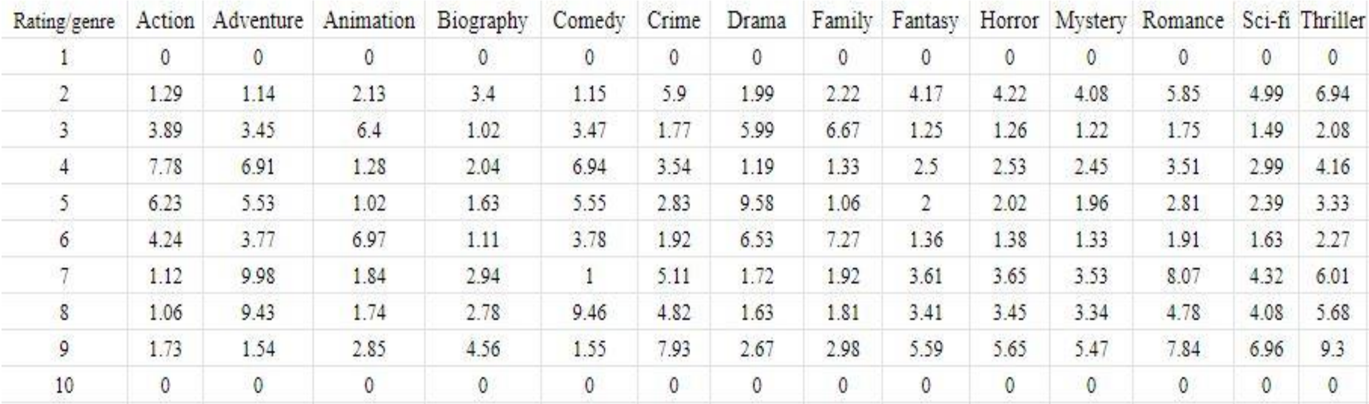
Table-1 : *Genres and Ratings Observed Frequencies*



Rating/genre	Action	Adventure	Animation	Biography	Comedy	Crime	Drama	Family	Fantasy	Horror	Mystery	Romance	Sci-fi	Thriller
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	1	1	0	1	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	2	0	0	1	0	0	1	1
5	6	5	0	0	4	0	8	1	2	6	1	4	4	7
6	57	45	2	0	45	16	41	7	22	27	9	17	19	20
7	129	88	12	19	116	54	141	25	37	43	40	61	45	56
8	79	99	26	46	78	53	208	14	27	16	32	44	35	61
9	13	16	7	9	11	7	40	2	4	0	8	3	6	8
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
total	286	254	47	75	255	130	440	49	92	93	90	129	110	153

Expected frequencies are calculated as *Count(Genres) x Count(Ratings)/n*

Table-2 : *Genres and Ratings - Expected frequencies*



Rating/genre	Action	Adventure	Animation	Biography	Comedy	Crime	Drama	Family	Fantasy	Horror	Mystery	Romance	Sci-fi	Thriller
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.29	1.14	2.13	3.4	1.15	5.9	1.99	2.22	4.17	4.22	4.08	5.85	4.99	6.94
3	3.89	3.45	6.4	1.02	3.47	1.77	5.99	6.67	1.25	1.26	1.22	1.75	1.49	2.08
4	7.78	6.91	1.28	2.04	6.94	3.54	1.19	1.33	2.5	2.53	2.45	3.51	2.99	4.16
5	6.23	5.53	1.02	1.63	5.55	2.83	9.58	1.06	2	2.02	1.96	2.81	2.39	3.33
6	4.24	3.77	6.97	1.11	3.78	1.92	6.53	7.27	1.36	1.38	1.33	1.91	1.63	2.27
7	1.12	9.98	1.84	2.94	1	5.11	1.72	1.92	3.61	3.65	3.53	8.07	4.32	6.01
8	1.06	9.43	1.74	2.78	9.46	4.82	1.63	1.81	3.41	3.45	3.34	4.78	4.08	5.68
9	1.73	1.54	2.85	4.56	1.55	7.93	2.67	2.98	5.59	5.65	5.47	7.84	6.96	9.3
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0

From the above Table 1, by using Chi square analysis we get the critical value as 114.268 with a significance of 0.05% and statistic was observed to be 207.818. And the dof=91. Since statistic was observed to be higher than the critical value, we can reject the null hypothesis that genres and ratings are independent. Hence we can establish that genres and ratings are strongly correlated.

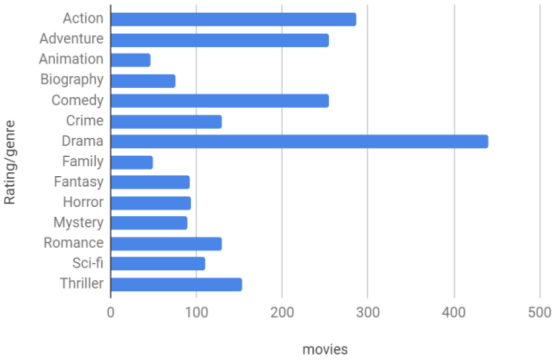


Fig 2 : Movies by genres

One of the problem we encountered while comparing the genres with their ratings was that several movies had more than one genre associated with them. So, we considered the total instances or occurrences of each genre in the dataset and performed the chi-square analysis. And since many movies had several genres associated with them, it also led us to the idea of finding out which pairs of genres were strongly correlated or say, which combination yielded a better result in the rating. Hence, we calculated the pairwise correlation coefficient for some of the pairs and the results were as follows:

i) Romance vs comedy:

Rating/genres	Comedy	Romance
1	0	0
2	1	0
3	0	0
4	0	0
5	4	4
6	45	17
7	116	61
8	78	44
9	11	3
10	0	0
Total	255	129

On performing the correlational analysis on the ratings of comedy and romance, the correlation coefficient was found to be 0.992748874 which was the highest among all the pairs. Thereby, we can confirm that romance and comedy are strongly correlated and could give a better result for a movie.

ii) Action vs Adventure:

Rating/genres	Action	Adventure
1	0	0
2	0	0
3	1	1
4	1	0
5	6	5
6	57	45
7	129	88
8	79	99
9	13	16
10	0	0
total	286	254

On performing the correlational analysis on the ratings of action and adventure, the correlation coefficient was found to be 0.943174, which suggests that action and adventure is also a better combination of genres with 94% correlation.

iii) Animation vs. Family

Ratings /Genres	Animation	Family
1	0	0
2	0	0
3	0	0
4	0	0
5	0	1
6	2	7
7	12	25
8	26	14
9	7	2
10	0	0
Total	47	49

On performing the correlational analysis on the ratings of animation and family movies, the correlation coefficient was found to be 0.717547, which suggests that animation and family genres are less correlated and do not go that well with each other.

Table-3. *Chi square analysis between Actors and Ratings*

Ratings/Actors	Christian bale	Tom cruise	Kate hudson	Total
<=6	2	1	3	6
>=6	11	8	0	19
Total	13	9	3	2

From the above table, by using Chi square analysis we get the critical value as 5.991 with a significance of 0.05% and statistic was observed to be 10.849, with dof=2. Since statistic was observed to be higher than the critical value, we can reject the null hypothesis that actors and their movie ratings are independent. Hence, we can establish that actors and their movie ratings are related.

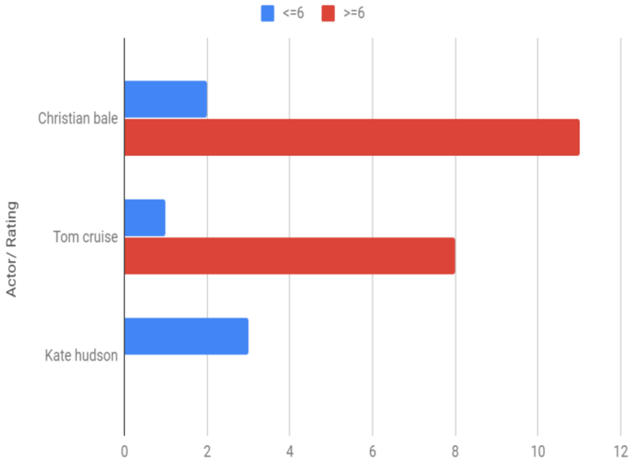


Fig 3 : Actor vs Ratings

This means that actors can be predicted to increase the ratings values of their movies. As shown by the bar graph above, most of movies of the famous actors, Christian bale and Tom cruise have a rating greater than 6. And the most of the movies of a lesser known actor like Kate Hudson have a rating lesser 6.

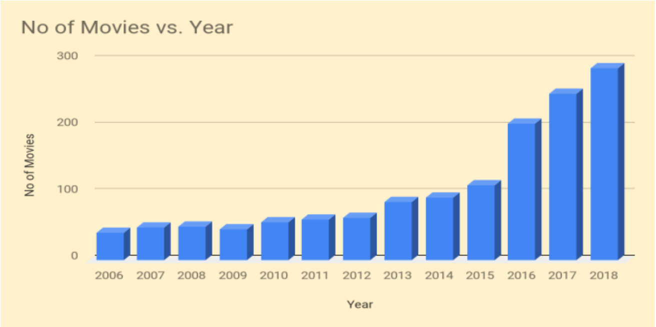


Fig 4 : No of Movies per year

VII.CONCLUSION

In our work, we have used a statistical model called Pearson's correlation analysis for finding out the correlation between various attributes related to a movie. Through the analysis, we could conclude that genres play a very important role in the success of a movie. And also that certain combinations of genres for a movie could yield better results in the success of a movie than the other combinations. We could also conclude that the actors who acted in the movie also played a part in the ratings obtained for a movie. Our work and results can be used to predict success or failure of upcoming movies by using prediction techniques like decision trees.

VI. REFERENCES

[1]. Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles, "Movie Success Prediction Using Data Mining".

[2]. Jiawei Han, Jian Pei, and Micheline Kamber. "Data Mining Concepts and Techniques", 2012.

[3]. M. Saraee, S. White, and J. Eccleston. "A data mining approach to analysis and prediction of movie ratings", 2004.

[4]. Nithin VR, Pranav My, Sarath Babu PBz, Lijiya Az, "Predicting Movie Success Based on IMDb Data",2017.

[5]. K Meenakshi1, G Maragatham2, Neha Agarwal3 and Ishitha Ghosh4, "A Data mining

Technique for Analyzing and Predicting the success of Movie", 2018.

[6]. "Success Prediction of Films at Box Office Using Machine Learning" by Parag Ahivale, Omkar Acharya.

[7]. "Predicting box-office success of motion pictures with neural networks" by Ramesh Sharda, Dursun Delen, 2015.

Cite this article as :

Shalini Mounika Maturu, Tamma Eekshitha, Talla Ramya,Tulluru Thanmai Durga, "Multi-genre Movie Data Analysis Using Pearson's Correlation", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 881-887, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952102
Journal URL : http://ijsrcseit.com/CSEIT1952102