Box Office Prediction

Lingtong Kong
10439934

Abstract— In the United States of America, 1000s of films are released ever year. Cinema in America is a multi-billion dollar industry where even individual films earn over a billion dollars. In recent years worldwide movies box office has also growing up very quickly.This research helps investors associated with this business for avoiding investment risks. The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from TMDb, IMDb and Wikipedia. Using Support Vector Classification (SVC), Decision tree predicts a movie box office profit based on some features. This paper shows Decision gives an accuracy of 82.7% for all features while Support Vector Classification (SVC) gives an accuracy of 85.0% for all features. Moreover, we figure out that budget, TMDb rates and run time are the most important features which play a vital role while predicting a movie's box-office success.

Keywords— movie industry; machine learning; support vector machine; neural network; decision tree; factorization machines; sentiment analysis; sparse data; data mining

I.Background and Related Works

1-1. Background

The movie industry is a massive sector for investment but larger business sectors have more complexity, and it is hard to choose how to invest. Furthermore, significant investments come with more significant risks. As movie industry is growing too fast day by day, there are now a considerable amount of data available on the internet, which makes it an exciting field for data analysis. Predicting a movie's box office success is a very complicated task to do. In the United States of America, 1000s of films are released ever year. Cinema in America is a multi-billion dollar industry where even individual films earn over a billion dollars. In recent years worldwide movies box office has also growing up very quickly. In today's world, the film industry is becoming more and more competitive. According to IMDB, the average number of films has been produced every year is 2577. It is impossible for all of those movies to survive in this fierce competitive market. All the film companies are exploring the way to increase their movie box office gross by adjusting the cast, changing director, increasing budget and etc. Many factors may influence the audience's liking and movie box office gross, like the director, cast, genre and budget. I will use all the possible factors and data in the history to build the model to predict the movie box office gross in a preferable accuracy. Therefore, they can optimize the predicted movie box office gross by adjusting the plan. My project may help film investor to improve the profit and lower the risks. Fig 1 shows the total yearly box office gross from 2000-2018.

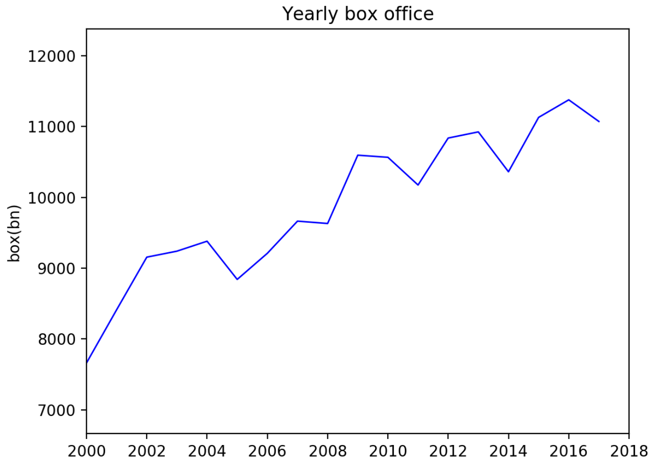


Fig 1 TOTAL YEARLY BOX OFFICE FROM 2000-2018.

1-2. Related Works

I have found two related works, the first is “A Machine Learning Approach to Predict Movie Box-Office Success”[1], and the author is Nahid Quader, Dipankar Chaki, Md. Osman Gani and Md. Haider Ali. They analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. And use Support Vector Machine (SVM), Neural Network and Natural Language Processing the system predicts a movie box office profit based on some pre-released features and post-released features. Their paper shows Neural Network and SVM both have a good accuracy of pre-released features and all features.

The second is “Movie Box Office Prediction System”[2] by Jiayong Lin. The author use Support Vector Machine to predict the box office. This research use the data form TMDB and analysis only USA movies.

II. Data Collection, Modification and Pruning

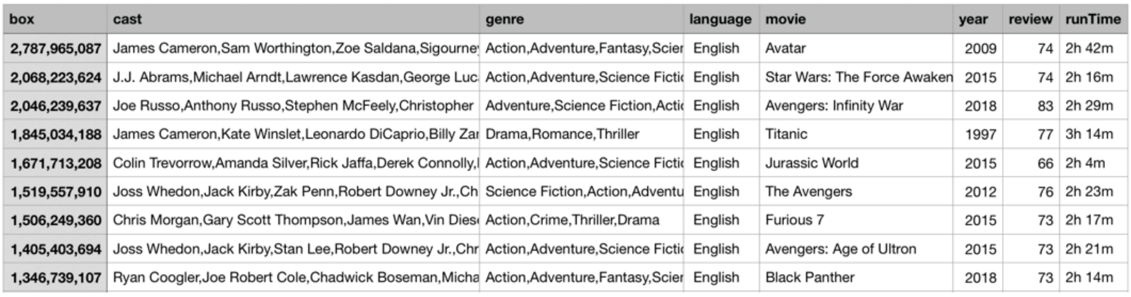
This section describes different phases of data preparation. This section include some steps such as data acquisition, data cleaning, feature extraction, data integration, and transformation.

2-1.Data Collection

I made a scrapy spider to collection datas. I get the movies' information from TMDB. The dataset contains more than 10000 movies released in between 1985 to 2018, data include movie Rating, gross, genre, language, famous stars and directors, year and runtime.

2-2. Data Cleaning

Unfortunately, Many data lack of gross. Because gross is the target values so I removed those information which lack of box office. And after that I got 6673 datas. Then we recognize that there are many movies which do not have all data attributes available. Most of the movies do not have the budget available.For some movies, we get the budget from Box-Office Mojo, IMDb and Wikipedia but still have about 900 movies in our dataset budget was unavailable in all sources and few movies do not have most of the features . After eliminating those movies, we finally make our dataset with 5170 movies which have all information available. Table 1 shows the part of the data.



box	cast	genre	language	movie	year	review	runTime
2,787,965,087	James Cameron,Sam Worthington,Zoe Saldana,Sigourne	Action,Adventure,Fantasy,Scier	English	Avatar	2009	74	2h 42m
2,068,223,624	J.J. Abrams,Michael Arndt,Lawrence Kasdan,George Luc	Action,Adventure,Science Fictic	English	Star Wars: The Force Awaken	2015	74	2h 16m
2,046,239,637	Joe Russo,Anthony Russo,Stephen McFeely,Christopher	Adventure,Science Fiction,Actic	English	Avengers: Infinity War	2018	83	2h 29m
1,845,034,188	James Cameron,Kate Winslet,Leonardo DiCaprio,Billy Zar	Drama,Romance,Thriller	English	Titanic	1997	77	3h 14m
1,671,713,208	Colin Trevorrow,Amanda Silver,Rick Jaffa,Derek Connolly,	Action,Adventure,Science Fictic	English	Jurassic World	2015	66	2h 4m
1,519,557,910	Joss Whedon,Jack Kirby,Zak Penn,Robert Downey Jr.,Ch	Science Fiction,Action,Adventu	English	The Avengers	2012	76	2h 23m
1,506,249,360	Chris Morgan,Gary Scott Thompson,James Wan,Vin Dies	Action,Crime,Thriller,Drama	English	Furious 7	2015	73	2h 17m
1,405,403,694	Joss Whedon,Jack Kirby,Stan Lee,Robert Downey Jr.,Chr	Action,Adventure,Science Fictic	English	Avengers: Age of Ultron	2015	73	2h 21m
1,346,739,107	Ryan Coogler,Joe Robert Cole,Chadwick Boseman,Micha	Action,Adventure,Fantasy,Scier	English	Black Panther	2018	73	2h 14m

Table 1 The part of the data.

2-3. Feature Transformation

From the table 2 we can see that many features like cast, genre are txt type, but many machine learning Algorithm can only use numbers so I got a famous actors/ actress and famous directors list from IMDb and calculate the famous casts number as one feature and recalculate the run time to minutes.Table 2 shows the summery of those dataset. From the data I notice that many movies has a lot of genres and language so I transform those features to one-hot feature. Finally I got 35 features.

Features	Mean	Median	Min	Max	Std.dve
Reviews	64.18224027858386	64	0	91	8.858623145564298
Casts Number	1.0437221899787192	1	0	7	1.2194114816546464
Run Time	115.18204681756626	106	26	338	29.65899470819
Budgets	33771181.53820855	20000000	6000	500000000	41729369.15220631

Table 2 Summery of Dataset.

III. EXPERIMENTAL RESULT AND EVALUATION

In this Project, I break the Movie Gross into 9 classes. Therefore, the problem becomes a multi-class classification problem. Table 3 shows the class of the gross and the Fig 2 shows the movie numbers of every class.

Class	1	2	3	4	5	6	7	8	9	10
Range(Million)	<1	1-10	10-20	20-40	40-65	65-100	100-150	150-200	200-250	>250

Table 3 Gross Class Table

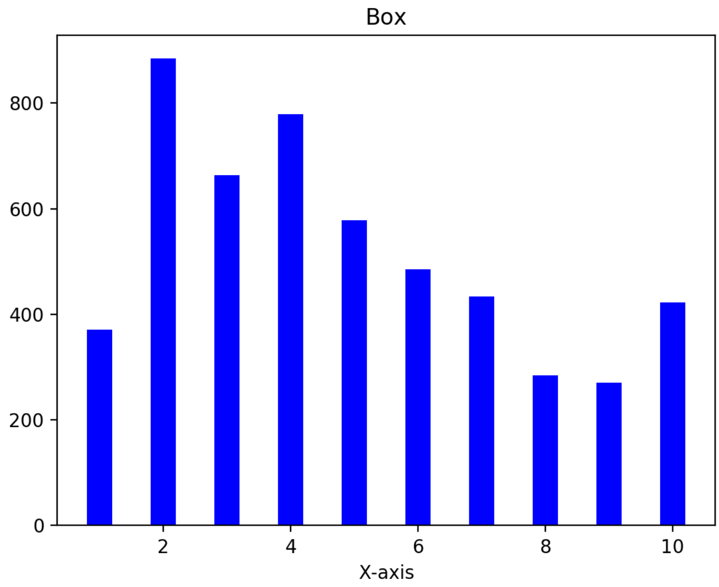


Fig. 2 Movie Numbers of Each Class

3-1. Decision Tree

I Implement 10-fold cross-validation. I separate the data in train data test data, the ratio of training to testing is 3:1. I chose Bagging decision tree classifier, extra trees classifier and random forest classifier. For the bagging classifier there has 200 trees in the forest, the number of samples to draw from X to train each base estimator is 0.5 and the number of features to draw from X to train each base estimator is 0.5. For the extra trees classifier the minimum number of samples required to split an internal node is 2, the number of features to consider when looking for the best split is the square root of n-features and I give it max 10 depth because if the tree in the forest is too deep will cause overfitting. The random forest classifier and the the extra trees classifier have the same parameters.Table 4 shows the accuracy of each classifier.

Classifier	Train Accuracy	Test Accuracy
Bagging Classifier	89.6%	66.2%
Extra Trees Classifier	80.7%	54.7%
Random Forest Classifier	93.9%	85.8%

Table. 4 The accuracy of each classifier

And for random forest classifier I also calculate the OOB Errors, the range of the tree numbers in the forest is 15 - 175. And I find that OBB Errors decrease significantly as the number increases as Fig 3.

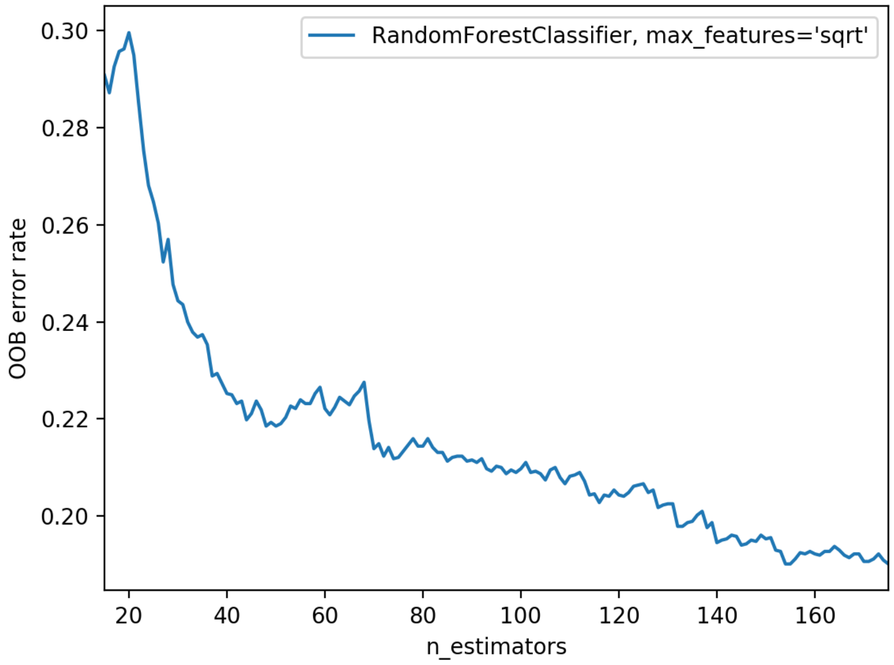


Fig. 3 OBB Error Decrease

And I also find the feature importance like Fig. 4, this picture show the total features' importance, the feature from 1-35 is 'Famous Cast', 'Genres: Fantasy', 'Year 1990-2000', 'Year before 1990', 'Year 2006-2012', 'Year 2000-2006', 'Year 2012-2018', 'Run Time', 'Movie rate', 'Genres: Adventure', 'Genres: Animation', 'Genres: Comedy', 'Genres: Thriller', 'Genres: ScienceFiction', 'Genres: Mystery', 'Genres: Horror', 'Genres: History', 'Genres: Romance', 'Genres: Drama', 'Genres:

Crime', 'Genres: Action', 'Genres: Documentary’, 'Genres: Music', 'Genres: Family', 'Genres: War', 'Genres: Western', 'Language: English', 'Language: Italian’, 'Language: Chinese', 'Language: Spanish;Castilian', 'Language: Japanese', 'Language: French', 'Language: Hindi’, 'Language: Other language', ‘Budgets’. From this picture we can find that Run Time, Movie rate and Budgets is the top three importance values. But if we put genres together like Fig 5 we can see that genres are more importance then run time and movie rate.

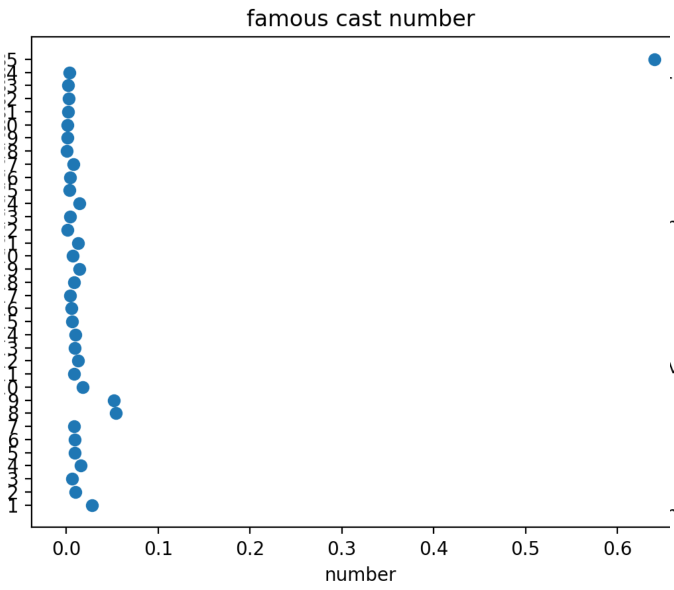


Fig. 4 35 features importance

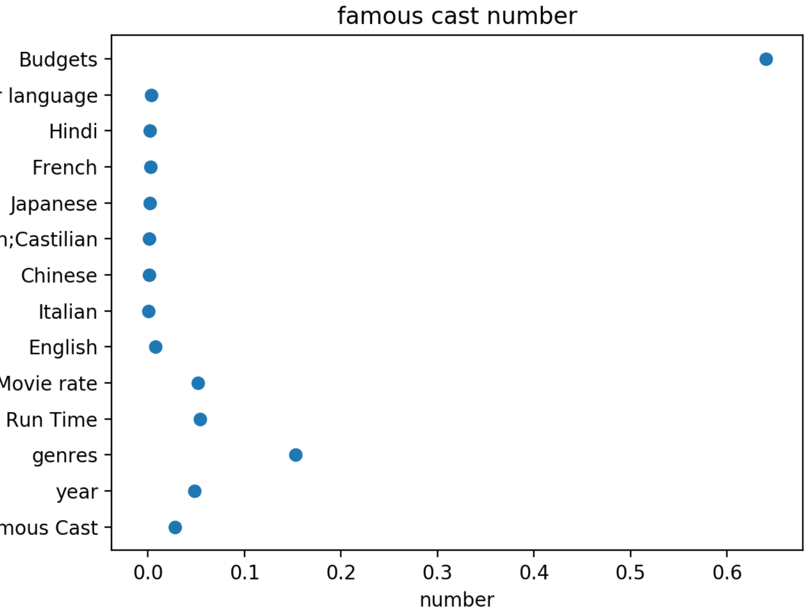


Fig. 5 features importance

3-2. Support Vector Classification

In SVC I use four kernel functions; the first one is the linear kernel, the second one is Gaussian radial basis kernel, sigmoid function and last one is 3-degree polynomial kernel. Kernels are useful for higher dimensional data as in practical life it is hard to calculate when we have more than two or three dimensions. We have 15 features as variables in this model and kernel function can work on infinite dimensional space. For this reason, we choose to apply all those kernel functions in our model. We use Scikit-Learn for the implementation of SVM. For linear kernel I set penalty parameter C of the error term is 0.1. In both Gaussian radial basis kernel and sigmoid kernel I set parameter C of the error term is 0.8, but the kernel coefficient are different, in Gaussian radial basis kernel is 0.01 and in sigmoid kernel I set it to 1 / (n_features * X.std()). In 3-degree polynomial kernel the penalty parameter C of the error term is 0.6 and kernel coefficient is 1 / n_features. And I use one-vs-rest decision function in all 4 training.

Kernel	Train Accuracy	Test Accuracy
Linear	88.3%	85.0%
RBF	58.5%	51.4%
Sigmoid	62.4%	64.1%
Polynomial	51.1%	37.8%

Table. 3 Accuracy of SVC

In Table 3, we can see the accuracy for SVC. Different kernels give us slightly different results. Among all those kernels Linear give a comparatively good result. Confusion Matrix for Linear is shown in Fig 6 for all features. When data are overlapping, SVM is unable to make hyperplanes properly. We know 2D plotting is a good way to visualize and understand the vector regions and data relations.

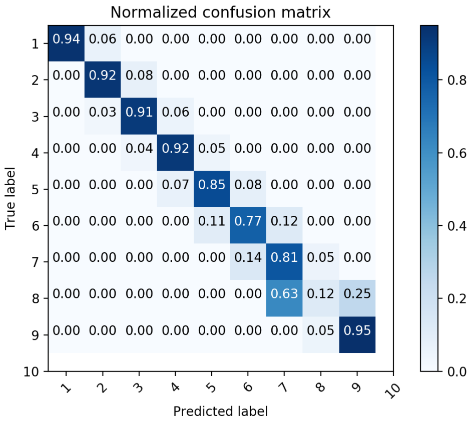


Fig. 6(1) Normalized Confusion Matrix

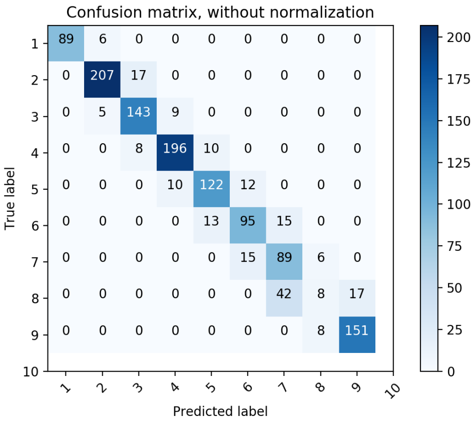


Fig. 6(2) Confusion Matrix, Without Normalized

IV. Analysis of The Results

From the above result we can see that both SVC and Decision tree can have a good accuracy, but we also find that the budget is too import to the result. In that case I want to remove the budget to find out what accuracy we can get unfortunately both SVC and Decision tree get a really low accuracy(both lower then 30%), so I break box office in two classes: more then 150million and less then 150million and I heard about another train module called Factorization Machines(FM)[3]. The FM use function to predict y

$$y(x) := \omega_0 + \sum_{i=1}^{n} \omega_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j$$

In the function ω_0 is global bias, ω_i models the strength of the i-th variable, $< v_i, v_j >$ models the interaction between the i-th and j-th variable.
In my project I gave $< v_i, v_j >$ 20 length and use use SGD and the loss function is: $-ln\sigma(y_1, y_2)$, σ is sigmoid function: $f(z) = \frac{1}{1 + e^{-z}}$. For every parameters the gradient is:

$$\frac{\partial y(x)}{\partial \theta} = \begin{cases} 1 & \theta = \omega_0; \\ x_l & \theta = \omega_l, l = 1,2,...,n; \\ x_l \sum\limits_{s=1;s\neq l}^{n} v_{sm}x_s & \theta = v_{lm}, l = 1,2,,n; m = 1,2,...,k, \end{cases}$$

After train 200 times we get train accuracy:75.4% and test accuracy: 75.3%.
But we only get two class if we want to break into more classes I think we need other information such as box office appeal of every stars and director, movies popular rate, Publisher level, Production Company level and Number of Screens.

V. CONCLUSION AND FUTURE WORKS

A movie success does not only depend on those features related to movies. The number of audience plays a vital role for a movie to become successful. Different countries have different population and different purchase power. Because the whole point is about viewers, the entire industry will make no sense if there is no audience to watch a movie. The number of tickets sold during a specific year can indicate the number of viewers of that year. And the purchase power should be considered by economic stability of a country. A country's GDP rate can be considered as a feature to know if there is financial stability during the period when a movie is released. And the year could be a really complex feature to analysis, because audience will change a lot over times and economic turn down or turn up over years. The general idea is during an economic depression, very few amount of audience will go to the theaters to enjoy movies, but on the other hand during the great depression the film industry increase about 700million dollars[4]. So these facts play a vital role in an ultimate success of a movie. So, for future work, we suggest considering these features. I do not box office appeal and sequel to a movie as features. Prediction of a sequel movie is terrible. Some movies gain a handsome amount of profit only for its previous sequel. And people from different countries will be interested in different movies genres that could be also difficult to consider and calculate.

I also think about only use poster and movies' storyline to predicted box office. I think we can get a lot of information from the poster such as stars director Production

Company and genres, and from genres we can also get some information like this. And different movies may face the audiences of different age, I should also consider the audience age in future work. And for future work, I want to predict the profit of the movies, because some movies may have a great box office but also has a great budget and this case may cause deficit.

REFERENCES

[1] Quader, Nahid & Gani, Md & Chaki, Dipankar & Ali, Md. (2018). A Machine Learning Approach to Predict Movie Box-Office Success. 10.1109/ICCITECHN.2017.8281839.

[2] Jiayong Lin. (2016). Movie Box Office Prediction System https://www.kaggle.com/tmdb/tmdb-movie-metadata/discussion/28576

[3] Steffen Rendle. (20 January 2011). Factorization Machines. 10.1109/ICDM.2010.127

[4] Robert S. McElvaine. (1993). The Great Depression: America, 1929-1941

[5] B. Pang and L. Lee, “Thumbs up? Sentiment classification using machine learning techniques,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, pp. 79–86.

[6] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[7] S. Gopinath, P. K. Chintagunta, and S. Venkataraman, “Blogs, Advertising, and Local-Market Movie Box Office Performance,” *Management Science*, vol. 59, no. 12, pp. 2635–2654, 2013.

[8] M. C. A. Mestyán, T. Yasseri, and J. Kertész, “Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data,” *PLoS ONE*, vol. 8, no. 8, 2013.

[9] J. S. Simonoff and I. R. Sparrow, “Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers,” *Chance*, vol. 13, no. 3, pp. 15–24, 2000.

[10] A. Chen, “Forecasting gross revenues at the movie box office,” *Working paper, University of Washington, Seattle, WA*, June 2002.

[11] M. S. Sawhney and J. Eliashberg, “A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures,” *Marketing Science*, vol. 15, no. 2, pp. 113–131, 1996.

[13] P. Chaovalit and L. Zhou, “Movie review mining: a comparison between supervised and unsupervised classification approaches,” in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2005.

[14] A. Sivasantoshreddy, P. Kasat, and A. Jain, “Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining,” *International Journal of Computer Applications*, vol. 56, no. 1, pp. 1–5, 2012.

[15] R. Sharda and D. Delen, “Predicting box-office success of motion pictures with neural networks,” *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006.

[16] W. Zhang and S. Skiena, “Improving Movie Gross Prediction through News Analysis,” 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009.

[17] M.H Latif, H. Afzal "Prediction of Movies popularity Using Machine Learning Techniques," National University of Sciences and technology, H-12, ISB, Pakistan.