ABSTRACT:

The history of motion pictures started way back in 1890 and this roots travelled to India in 1910. Earlier movies involve only sets to display but movies nowadays involve real scenarios which is risky endeavour. India ranks first in terms of annual film output and Bollywood is the largest and most powerful branch and controls Indian cinema.

The success of a film depends on the box office revenues. Out of 250 bollywood movies released in 2016, less than 10 movies have crossed revenue mark of ₹100 million($1.56 million). Major of the films failed to regain the money which is invested in motion pictures. This paper deals with the prediction of success or failure of movie to prevent loss.

In this paper, we used linear regression to forecast the movie revenues and implemented different techniques to increase accuracy of model. In addition to these methods, Sentiment Analysis and Analytical Hierarchical Process(AHP) also used for giving scores to reviews and giving values to nominal data like ranking actors, directors, genre and production respectively.

INTRODUCTION:

1.1 The Indian motion picture industry:

Indian motion picture industry is the most prolific film industry in the world that has been largest movie producer for the last few years with staggering 1200 films every year in 23 languages.There are over 2200 multiplexes in India and in 2016 more than 2.2 billion tickets were sold in india compared to 1.25 billion sold in second placed China.

Bollywood,world famous hindi film industry is the largest film producer in India.bollywood acquired its name based on the one of the most important cities in the global film industry,Bombay. It is this city which is the center and the birthplace of bombay, the multi-million dollar Indian film industry.Bollywood represents 43% of the net box office while Telugu and Tamil cinema represents 36% and other regional cinema constitutes 21% as of 2014.Bollywood is also one of the largest film industries in the world in terms of the number of people employed.

The value of the bollywood could reach upto $2.89 billion(INR 193 billion) from its current $2.32 billion(INR 155 billion) ,according to the report, “Digitization & Mobility: Next Frontier of Growth for M&E”,from accountancy firm Deloitte and The Associated Chambers of Commerce and Industry of India.(ASSOCHAM).

Data Collection:

The initial dataset used in this project is extracted from IMDb website. We extracted only the bollywood movies which were released in between the years 2009 and 2016. Additional data is extracted from the website bollywoodhungama. Title of movie, year of release, rating, reviews for each rating, metascore, number of votes for rating, runtime of movie, description, directors, actors, genre, production company, central board of film certification are extracted from IMDb and release dates, opening day collection, opening weekend collection and one week collection of each movie are extracted from bollywoodhungama. We extracted total of 1745 movies from IMDb, 1084 movies from bollywoodhungama and 8735 movie reviews from IMDb.

Tools used:

The method of extracting data from websites is known as web scraping. Rvest is a package present in R language that is used to scrape data from html web pages(similar to beautifulsoup in Python). Web scraping packages can access the world wide web using Hypertext Transfer Protocol(HTTP) or through a web browser. Web pages are build using HTML or XHTML(text based markup languages). In some websites, like twitter and facebook, webpages are build in such a way that the data in the web pages cannot be scraped directly. They provide Application Program Interfaces(API) keys to each of the users so that security of data is maintained. Data from IMDb, Bollywoodhungama and TOI websites can be directly scraped without using API. To extract particular type of data, HTML/CSS tags are required. SelectorGadget, an open source tool present in chrome extension, that makes CSS selector generation and discovery in a much easier way from complex websites.

Data Preprocessing:

The Data thus obtained is highly likely to have missing,inconsistent and noisy data due to its big size and because of scraping from many heterogeneous sites.The main sites from which the data is scraped are IMDb,Times of India and Bollywood Hungama.The main problem with the data scraped was missing fields.This was overcome by obtaining the missing data from alternative sites. Similarly inaccurate and corrupt records from the database were detected and were either removed or corrected.

Data Integration and Transformation:

The Data scraped from these websites is merged and stored in SQLite database.This integrated data is then transformed into coherent forms appropriate so that further processes(Regression) would be easier and efficient.The data collected has both numerical and nominal attributes.Since we need all the attributes to be numerical, nominal attributes are to be converted into equivalent numerical values.Here,(instead of/along with) using a measure of central tendency of Box office revenue to convert corresponding nominal attributes to numerical we used AHP(Analytic Heuristic Process) to calculate the numerical values of the nominal attributes.

Collected data:

NOMINAL ATTRIBUTES: Directors, Actors, Genre, Production-House, Certificate, reviews, movie description, release date.

NUMERICAL ATTRIBUTES: Rating, Critics_Rating, Votes, Gross_Collection, Opening_Day_Collection, Opening_Weekend_Collection, First_Week_Collection, Budget, metascore, time of movie, votes.

Used data for AHP:

NOMINAL ATTRIBUTES: Directors, Actors, Genre, Production-House.
NUMERICAL ATTRIBUTES: Rating, Critics_Rating, Votes, Gross_Collection, Budget, metascore, time of movie, votes.

For regression all the attributes have to be numerical but the dataset consists of both nominal and numerical attributes . Hence we used AHP to convert corresponding nominal attributes to numerical.

Converting Nominal to Numerical Data:

The filtered data thus obtained has both numerical and characters. Nominal parameters like actors, directors, genre and production company cannot be used for the prediction model. But these parameters play important parameters in building a model. So, we have to assign a value to the above parameters. These values are assigned based on the previous numerical data. The weights corresponding to the numerical parameters is assigned using the method Analytical Hierarchy Process(AHP). Using these weights, we calculate numerical values for nominal data. Before assigning numerical values to nominal data, we normalized all the numerical data to the scale 0-10 so that bias can be reduced. For example, value of Ratings will be in scale of 10 but value of budget and gross will be in millions which makes value of ratings as negligible.

Selecting feature Subset:

Model Generation:

Supervised learning technique is adopted for this Study. We used //three// models to predict the revenue and we will compare the performance of the different methods.

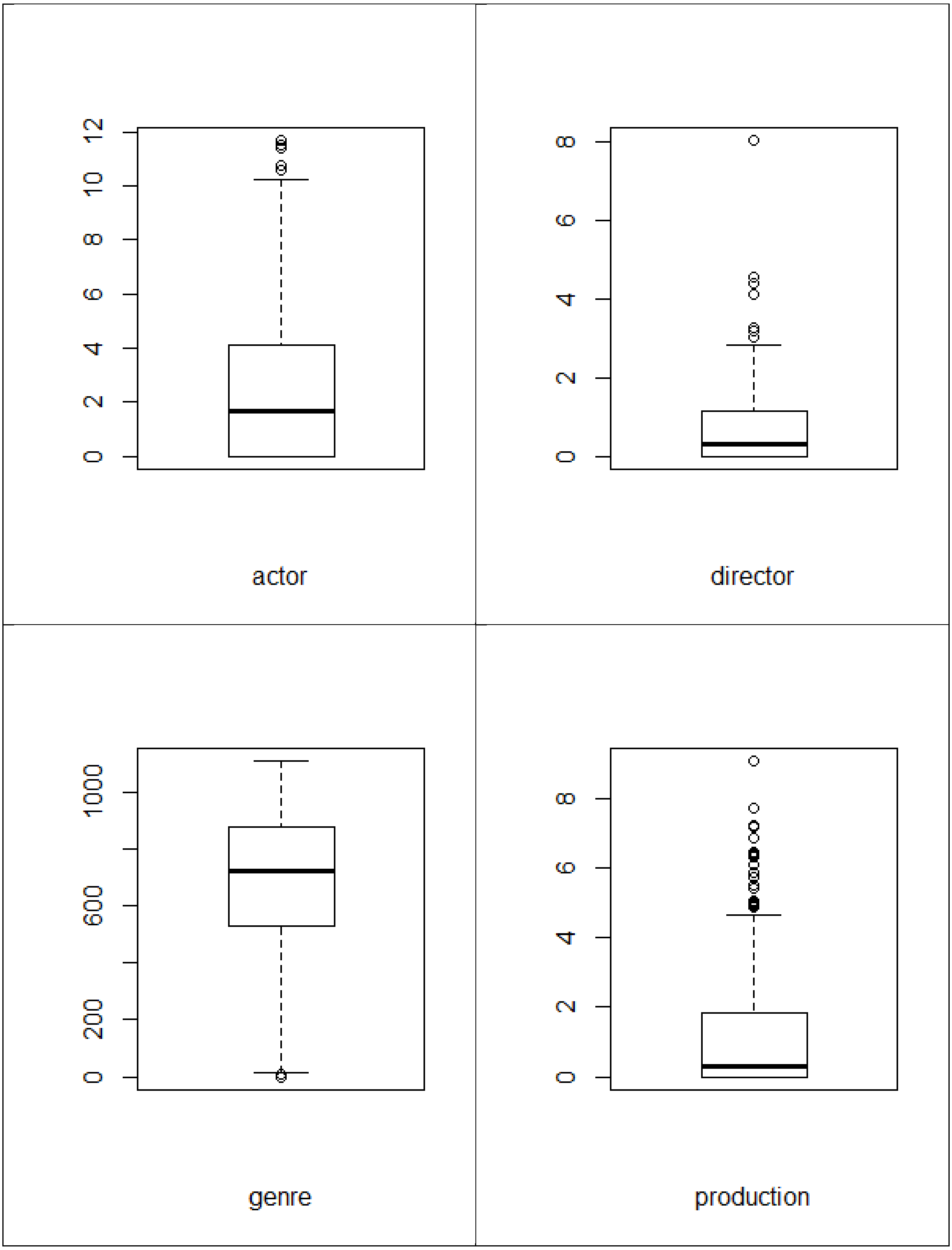
Linear Regression Model :

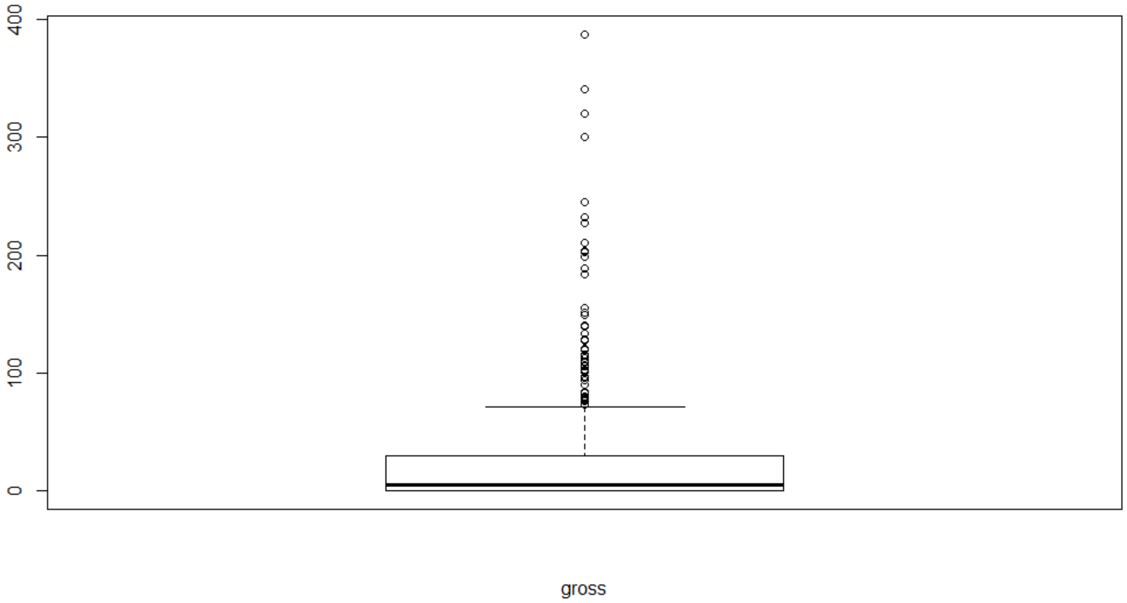
The first model is standard least squares linear regression. Stochastic gradient descent is used to do this. Once we have trained set of feature weights, we could then generate gross revenue predictions as follows:

$$Gross=0+1*F1+2*F2+.....+n*Fn \qquad (1)$$

Where i are the weights, Fi are the features, and n is the number of features.

The model is generated using the input variables directors, actors, genre, production-house which are generated using AHP. Other parameters like ratings are not considered since ratings are not generated before the release of the movie. Due to the lack of data, budget is also not considered.


0
2
4
6
8
10
12
actor
0
2
4
6
8
director
0
200
600
1000
genre
0
2
4
6
8
production



Linear regression is applied and adjusted R-square turned out to be low. Applying box-plot for each variable, outliers turned out to be high in number for production and gross. Accuracy of the model increased when the outlier values are replaced with nearest value which is in range. The values which are obtained above are not satisfactory. Box cox transformation is applied on output variable which improved the value of adjusted R-square from 0.54 to 0.64. Box cox transformation converts non normal dependent variables to normal variable which is an assumption for applying linear regression.

		R-square	Adj. R-square	AIC	BIC
1	Linear regression	.4591	.4557	6225.161	6251.485
2	Linear regression with adjusting outliers	.5435	.5406	4923.272	4949.956
3	Linear regression with box cox transformation and adjusting outlier	.6349	.6326	2909.363	2936.047
4	Linear regression with box cox transformation	.6409	.6386	3257.557	3284.24

The value of AIC and BIC is high since genre and gross are in scale of thousands. By comparing values, models 3 and 4 are better when compared to models 1 and 2.

Conclusion:

After refining the model we found out that this multiple linear model with box cox transformation represents the feature of the movies more accurately. The accuracy of the model is not good enough but is still better than other previous studies and some results are better than that of some standard libraries and such studies. Model will be more accurate if we increase the training data set or increase the features of dataset like considering News feed and Social Network's data like twitter and facebook. Analysis could be done on this data and it could be added to the training set to improve the accuracy of the model. This models can be used for industrial purposes or online applications.

References:

Nithin VR,Pranav M, Sarath Babu PB, Lijiya A, "Predicting Movie Success Based on IMDB Data", International Journal of Business Intelligents, Vol. 3, Issue 2, 2014, pp. 34- 36.

M. Ghiassi , David Lio, Brian Moon, "Pre-production forecasting of movie revenues with a dynamic artificial neural network", Expert Systems with Applications, volume 42, Issue 6, 2015, Pages 3176-3193.

Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques", 3rd ed.MA:Elsevier, 2011, pp. 83-117.

Chen, X., Chen, Y., & Weinberg, C. B., "Learning about movies: the impact of movie release types on the nationwide box office", Journal of Cultural Economics, 2013, 37(3), 359–386.

Dhar, T., Sun, G., & Weinberg, C., "The long-term box office performance of sequel movies", Marketing Letters, 2012, 23, 13–29.