Movie Fortune Teller

Gurleen Singh Dhody
Indiana University
USA
gdhody@iu.edu

Madrina Thapa
Indiana University
USA
mthapa@iu.edu

Vinita Boolchandani
Indiana University
USA
vinitab@iu.edu

ABSTRACT

Movies are the most generic source of entertainment in the present world. Hence they also contribute towards one of the major revenue generating industries across the globe. In this paper we describe the development of a movie prediction model through a sequence of social media mining steps. The model primarily focuses on predicting success of a movie considering the effect of social media websites. In the current era of social media websites, user is given immense freedom to express their views on diverse topics. The repercussions of this freedom can be exploited to study the trend followed by performance of movies since the time their official trailers are released. This analysis has been conducted on data of movies from year 2008 to 2016.

CS CONCEPTS

• **Social Media Mining,** Regression Models, One Hot Encoding, Support Vector Machines.

KEYWORDS

Movie Success Prediction, web scraping, social media mining, data mining, Python

1 INTRODUCTION

In the last decade we have seen a revolution of social media websites. In this age of digitalization, we have been gifted with an access to plethora of resources and data online. This comes with its own disadvantages as our routine today leaves a trace of almost everything we do. This includes a track of our interactions, transactions, locations and much more. However, it has also empowered us in terms of actively sharing information, expressing opinions and staying abreast with the latest issues. Social Media forms an important pillar of this digital world. It is the largest platform for communication between people across the world. Social media websites can also be used for analysis of various serious issues like Image Popularity prediction using sentiment and context[1]. It has also been used for predicting the future[2], predicting the flow of information through networks[3], and also for predicting the viral content[4]. That said, Social Media has also been used to predict the success of a movie.

Before the advent of Social Media, movies were promoted using the traditional forms of media like newspapers, TV channels, or dedicated websites. There was no global platform for promotion or for the audience to express their opinion or access others' opinion about a movie. In those times, the success of movie was attributed to the amount of revenue it generates. With the ubiquitous use of social media in last couple of years, it has been rendered as a platform for movie promotion. However the effect of this promotion platform has not been effectively used to establish a correlation between pre-release features and post release success. Moreover the fact that there is no universally accepted model, instigated an urge amongst us to develop one such relationship.

RELATED WORK

In the last few years there has been a lot of analysis to study the success prediction of a movie. Some of such interesting works were studied by us as a part of our initial research. One such analysis was done by Márton Mestyán, Taha Yasseri , János Kertész[5] where they have used Big Data to bridge the gap between "real time monitoring" and "early predicting". They have build a predicate model to financial success of movies based on collective data of online users. In this study the authors have restricted their study to the effect of corresponding movie entry to Wikipedia which according to them is the online encyclopedia. This subsequently differs from our analysis as we have tried to encompass data from various websites to take into account a variety of features that might affect the success of a movie. Another work that we came across was *Determinants of Motion Picture Box Office and Profitability:An Interrelationship Approach*[6]. In this paper the authors have tried to study the interrelationship among a number of factors that affect the movie profitability. Their dataset is composed of 331 movies. Although this paper was a motivation for our work we differ substantially in terms of the features studied and the vastness of dataset that we have considered. Additionally our source of dataset is essentially social media websites which is another point of disparity.

2 MODEL FORMATION PROCESS

2.1 Data Extraction

In current times movies are one the most important source of entertainment as well as revenue generation. Before a movie is released its success depends on factors like the existing popularity of the Director, popularity of Actors, promotions, genre, time of release etc constituting one dataset. Once the movie is released its success can be studied in terms of the profit it makes or the rating it earns on various social media websites. Factors like critic rating, audience rating, box-office life and awards are a part of the intermediate dataset that are in some way interdependent and dynamically keep contributing to final success or failure of a movie. In order to study these factors in details we start with extracting this data from four websites namely Facebook, IMDB, Google Movies and TheMovieDB.

2.1.1 Brainstorming

The first step in extraction was brainstorming the concept "Success". Success can be in terms of awards, rating, box office life or revenue generated. Additionally, success can differ according to who it concerns, whether movie or its cast. Since the complete analysis is about predicting 'success', we started off with getting our concepts correct. In the rest of this paper, we would be attributing the term success, concerning a a movie, to the IMDB rating it is awarded.

2.1.2 Technicalities of Extraction

The data extraction from social media sites was perhaps the most challenging task. Any analysis gains clarity only as it proceeds. Since this was the first step, it was quite demanding to extract the relevant feature from four different websites while minimizing the unwanted data. We accomplished this using Web Scraping. In order to extract sufficient data for training and validation, top movies from year 2008 to 2016 were targeted because Facebook expanded and started business pages in late 2007. This was implemented using Facebook's API[9], MovieDB API[10], IMDB Pie [8] and IMDBPY[7] with Beautiful Soup, a python library for pulling out data from HTML files (JSON format). Beautiful Soup can work with a variety of parsers (html.parser and lxml have been used by us) and also provides methods for traversing and searching elements of the parse trees. This scraped data was written to CSV files using the cvs package in Python. These year wise generated CSV files were placed into separate folders locations defined globally in the code. All the information was fetched dynamically from the internet using http request. Hence the scripts are independent to be used later on any system.

The raw data was essentially in the form of strings and numbers representing features (columns) like Title, Overview, Production, Release Year, Revenue, Budget, Runtime, Tagline, imdb_id, Genres, and user rating to name a few. In addition to the CSV files we also extracted some .html files from Google Movies. For future use, a list of the academy award winning films was also written to which included features like Movie_Name, Year, Awards, Nominations and Best_movie. Each row in the CSV file represents an individual movie instance. Most of the extracted fields are listed in Table 1.

Title	score	Budget	Runtime
faces	fb likes	IMDB_review_helpful_metric	Revenue
genre	IMDB_review_rating	IMDB_review_sentiment_neg	IMDB_review_sentiment_pos
IMDB_review_neutral_metric	IMDB_review_sentiment_neg	overview_neg_score	certificate_G
certificate_Unrated	certificate_TV-14	certificate_A.G.	IMDB_review_sentiment_pos

Table 1: List of features extracted

These features were in raw form with a lot of unwanted information. A sample of this data has been placed in the Appendix table A. This data had to be cleaned and important features to be segregated. For the detailed list of features used please refer Table 2.

2.2 Data Preprocessing

The raw data extracted in its current form had to be cleaned before it could be used for training. This was done using *pandas*. The data had some features that were binary consisting of either of the two values. Such data was processed using one hot encoding. There were categorical features as well. These were converted into numbers to be used by our model. All the values of a particular label will be made as a new feature. If an instance has this value for the respective feature, the value entered will be 1, else it will be 0. The features that was highly categorical in nature was genre. For analysis of raw text features like user reviews, critics reviews and overview we decided to use Abstract Syntax Grammar and Natural Language Toolkit. All the positive key words like "good movie", "promising" or "hit" were segregated into positive reviews and others were grouped as negative reviews.

2.3 Feature Identification

The preprocessed data had a lot of features that seemed important at prima facie. Generally there are a lot of extraneous data that is obtained when a webpage is scraped. This data may not necessarily impact the dependent variables. Such features that do not hold any direct or indirect significance in results are discarded at this stage. In order to make a decision about which data were relevant to our model we resorted to pattern generation.

2.3.1 Pattern Generation

The *data* that is extracted from internet can be visualized in form of patterns to study the relationship among various features. This visualization turns the raw data into useful *information*. In later stages this information is transformed into *knowledge* about past patterns which helps us to predict future trends.

Among the list of available features our primary aim was to identify dependent variables(y) and independent variables(x). Another goals was to generate patterns that demonstrate how the later governs the behavior of the former. In order to achieve this we made use of a number of Python packages and collaborating platforms. The major ones have been listed below with their significance.

LIBRARY/PACKAGE	PURPOSE
matplotlib	For plotting the data.
numpy/pandas	For storing year wise dataset table format.
sklearn	Used for training the model we chose that is Multi linear regression.
nltk.corpus (natural language toolkit)	For removing stopwords and performing sentiment analysis
cv2	Detecting faces in Facebook display picture.

Among all the patterns generated here are a few important ones to develop a basic idea of what we obtained.
The plot in Figure 1 is between different genres and their respective movie count over the years. Most movies belonged to *Drama* with the highest count of 250 followed by *Action* and *Comedy.*

Another important relationship was between the production houses and the number of popular movies produced from year 2008 to 2016, depicted in Figure 2.

The maximum number of successful movies were produced by Warner Bros. We can consider that the production house has an impact on the success. If they are leading the chart

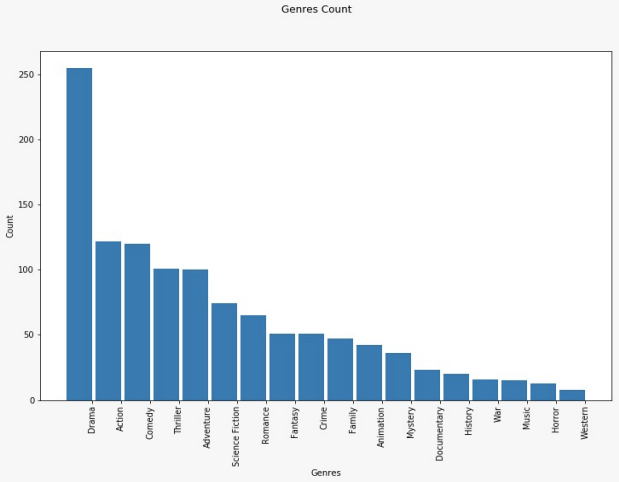

Figure 1: Different genres and count of movies in each genre

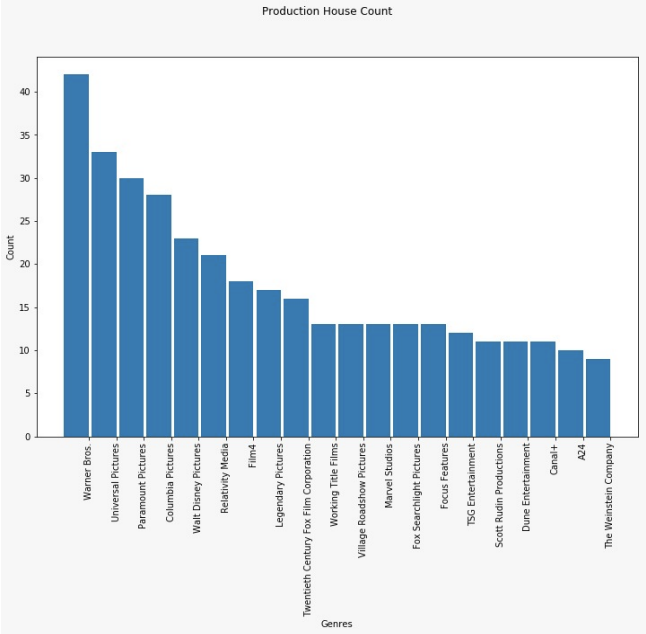

Figure 2: Top Production house and the number of movies produce by them in the year of 2008-2016

of successful movies the audience develop a proclivity towards them and look forward to their forthcoming release.

Figure 3, 4, 5 and 6 shows the movies released in the year of 2008, 2010, 2013 and 2016 respectively. For determining the Facebook likes, we have considered all the verified movie pages on Facebook and have taken the average of likes on all the pages to calculate the total Facebook likes. The red bars in the figures indicate that the movies have won Oscar in that particular year. It is evident from all the four above mentioned figures that the movies

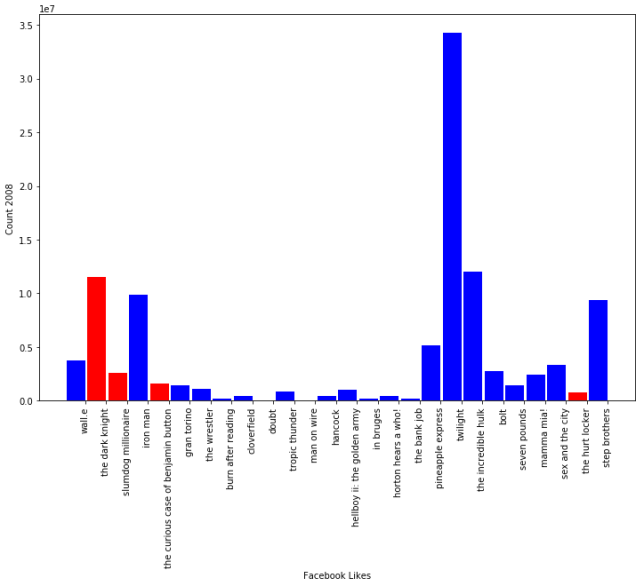

Figure 3: Facebook likes for movies released in the year of 2008

that has been the most popular on the Facebook has not won an Oscar. Whereas the outlier was seen in 2013, the movies 'Frozen' was the most popular on Facebook and also won an Oscar award.

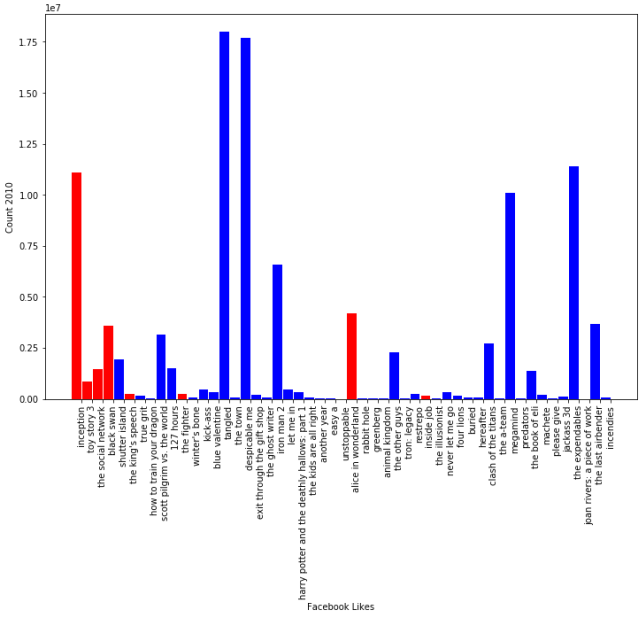

Figure 4: Facebook likes for movies released in the year of 2010

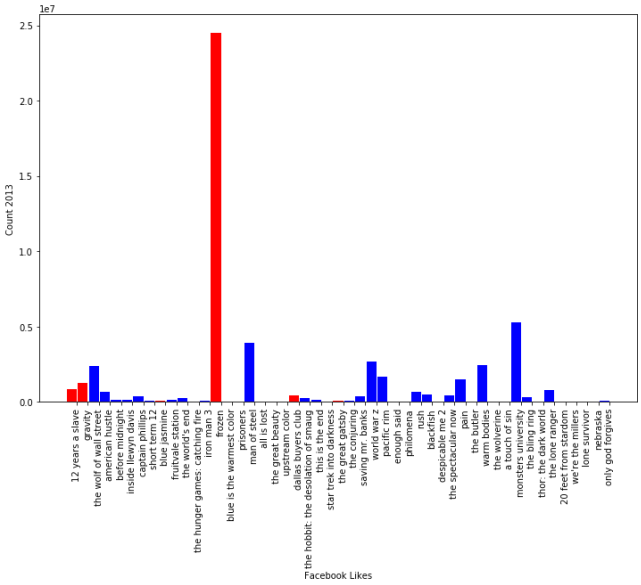

Figure 5: Facebook likes for movies released in the year of 2013

In the year 2013, there is another important thing to notice. Most popular movie on Facebook won an Oscar and there was a huge difference in the number of likes 'Frozen' had in compare to the other Oscar winning movies in the same year.

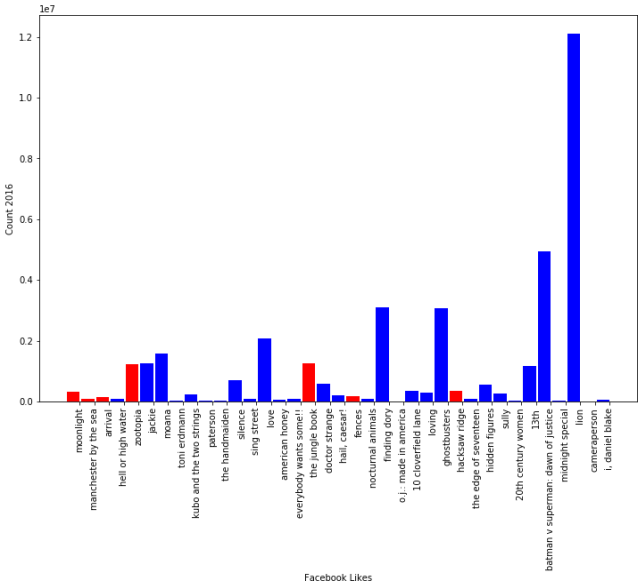

Figure 6: Facebook likes for movies released in the year of 2016

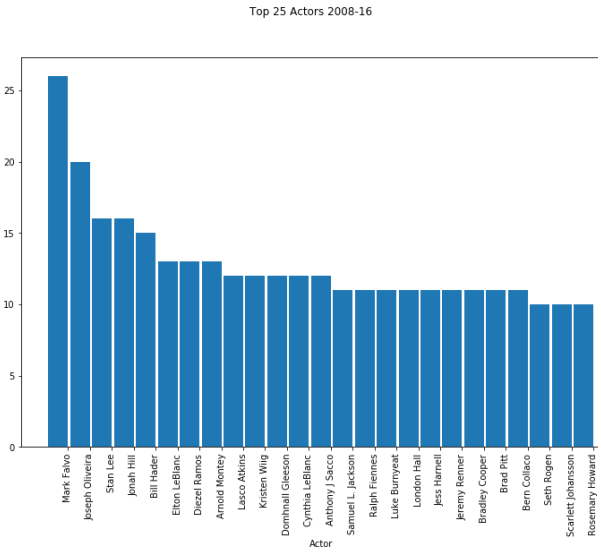

Figure 7: Top 25 Actors from 2008 - 2016

The top 25 Actors in 2008 - 2016 as shown in the Figure 7 shows that the Actors who did more movies are in the top of the lists are mostly not the lead actors.

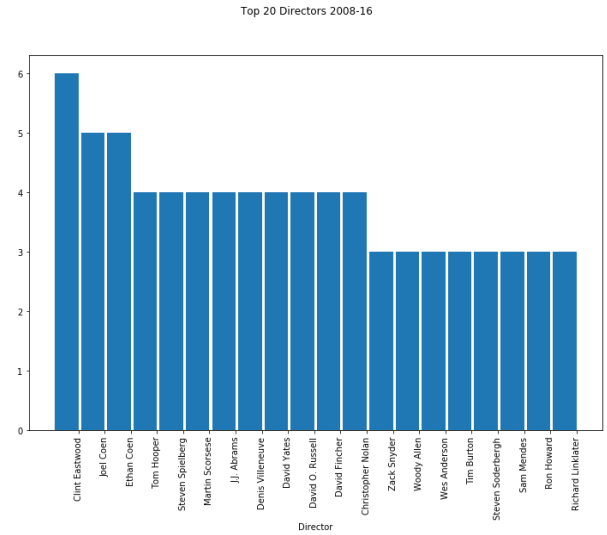

Figure 8: Top 20 Directors from 2008 - 2016

From Figure 8, top 20 Directors from the year 2008 - 2016 have directed more or less the same number of movies, as the difference is highest and the lowest is not more than 3. It is surprising to notice that there were no female directors in the list of top 20 directors from the year 2008 - 2016.

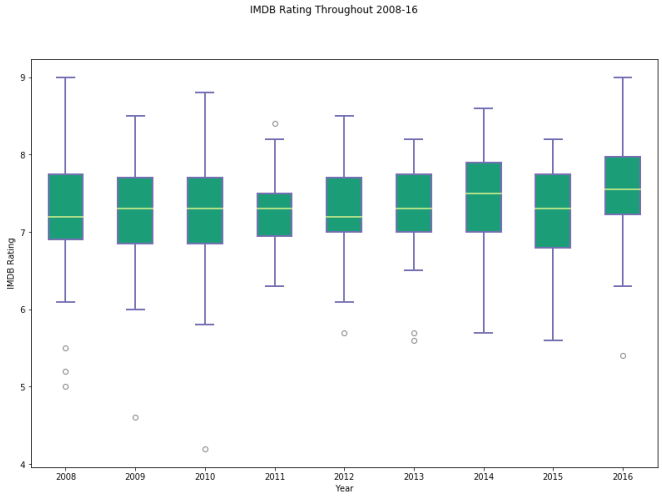

Figure 9: IMDB rating from 2008 - 2016

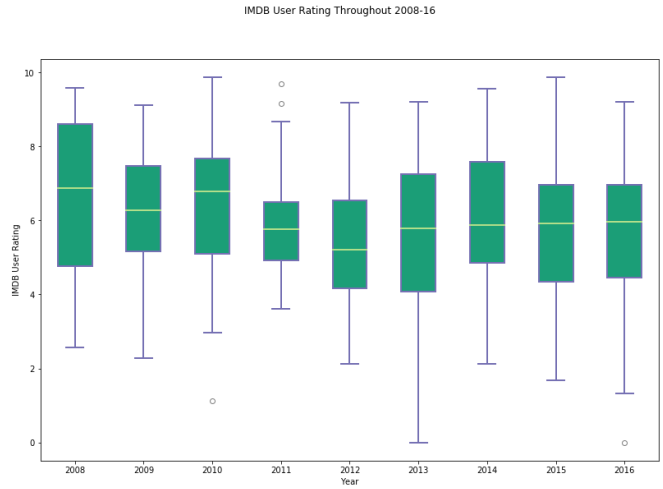

Figure 10: IMDB User rating from 2008 - 2016

We fetched the top 25 reviews listed in IMDB for each movie from the year 2008 - 2016 and considered the ratings given by the reviewers for them. The box plot in the Figure 10, shows the IMDB user rating derived from the top 25 reviews for each movie from the year 2008 - 2016. On comparing IMDB rating and IMDB user rating, shown in the figure 9 and 10 respectively, it is clear that IMDB rating and IMDB user rating has weak correlation. As the IMDB user ratings box plots are spread over and are more symmetric than compared to the IMBD rating plot.

Features	Description	Analysis/ Model
title	Title of the movie	Analysis
score	IMDB score of a movie	Analysis/Model
IMDB_review_helpful_metric (x/y)	Average ratio between users who found the review helpful and the user who viewed the review	Model
IMDB_review_neutral_metric ((y-x)/x)	Average ration between user who didn't react towards the review and users who found it helpful	Model
IMDB_review_rating	Cumulative rating based on IMDB user review	Model
IMDB_review_sentiment_neg	Cumulative negative score of IMDB movie user review	Model
IMDB_review_sentiment_pos	Cumulative positive score of IMDB movie user review	Model
budget	Total budget of the movie	Model
faces	Total number of faces detected in Facebook display picture	Analysis/Model
fb_likes	Total number	Analysis/Model
overview_neg_score	Cumulative negative score of movie overview	Model
overview_pos_score	Cumulative positive score of movie overview	Model
revenue	Income made by the movie	Model
runtime	Length of the movie in minutes	Model
Action	Genre of the movie	Analysis/Model
Adventure	Genre of the movie	Analysis/Model
Animation	Genre of the movie	Analysis/Model
Comedy	Genre of the movie	Analysis/Model
Crime	Genre of the movie	Analysis/Model
Documentary	Genre of the movie	Analysis/Model
Drama	Genre of the movie	Analysis/Model
Family	Genre of the movie	Analysis/Model
Fantasy	Genre of the movie	Analysis/Model
History	Genre of the movie	Analysis/Model
Horror	Genre of the movie	Analysis/Model
Music	Genre of the movie	Analysis/Model
Mystery	Genre of the movie	Analysis/Model
Romance	Genre of the movie	Analysis/Model
Science Fiction	Genre of the movie	Analysis/Model
Thriller	Genre of the movie	Analysis/Model
War	Genre of the movie	Analysis/Model
Western	Genre of the movie	Analysis/Model
certificate_*	All the available certificates for the movie(* here symbolizes that different kinds of certificates exist)	Analysis/Model

Table 2: Detailed list of features extracted

We considered the count of faces on the display picture of the Facebook movie page and used it as a feature. We used *opencv* for face detection.

We performed sentiment analysis on movie overview and IMDB user reviews. For sentiment analysis we used ntlk corpus that had the following modules - *sentiwordnet, stopwords, regexptokenizer, wordnetlemmatizer, wordnet.* Our approach for analyzing negative and positive sentiments are as follows.

- Stopwords were removed from user review.
- All numerical and special characters were removed from user review.
- All root were extracted.
- Positive and negative sentiment for each remaining words in the user review were summed and averaged.

2.4 Model used for Training

After preprocessing and feature selection we were at the stage of training our data. Our goal in this study was to develop a model that would be able to predict the success

of a movie from given features. Hence a supervised learning model would be appropriate for training the data. This boiled down to choosing either of Linear Regression Model and Multi Linear Regression Model. Linear Regression Models are usually implemented to predict single independent variables. Whereas Multi Linear Regression Models are used to predict when there are more than one independent variables. As we were anticipating to predict IMDb user rating, to show successful a movie will be in a social platform. We chose Multi Linear Regression Model for training our data. The output that our model estimates is the success of a movie in terms of its IMDB user rating.

Table 2 has a detailed list of the features we extracted with a small description about it along with the information about whether we used it for analysis, modeling or both.

The plan outlined in this document would help us study the relation between the movies and their performance after the release. As we are extracting the real time data, after cleaning and preprocessing it will be integrated with the data of other websites. After integration we would have the most relevant data that would supposedly yield us the relationship that we are seeking with this study. This data would then be trained into a model performing a k-fold validation. After the validation the model would be run on the test data and if successful the resulting model would help us predict the success of a movie as output given its independent input variables.

3 RESULTS

We had collected 450 movies between 2008 - 2016 during data collection phase and approximately 75% of the collected data complied with all the constraints of our multi linear regression model. The complete dataset was partitioned into two groups for training (90%) and testing (10%), where best hyper parameters were determined through 10-fold Cross Validation. For Multi Linear Regression we used Elastic net and Ridge regression. The R2 score is 0.65 and 0.60 respectively.

The result of our Multi linear regression model was plotted against the ground truth value. Please refer Figure 11. The blue dots here represent the true IMDB User rating and red cross represent IMDB user rating predicted by our trained Multi Linear regression Model.

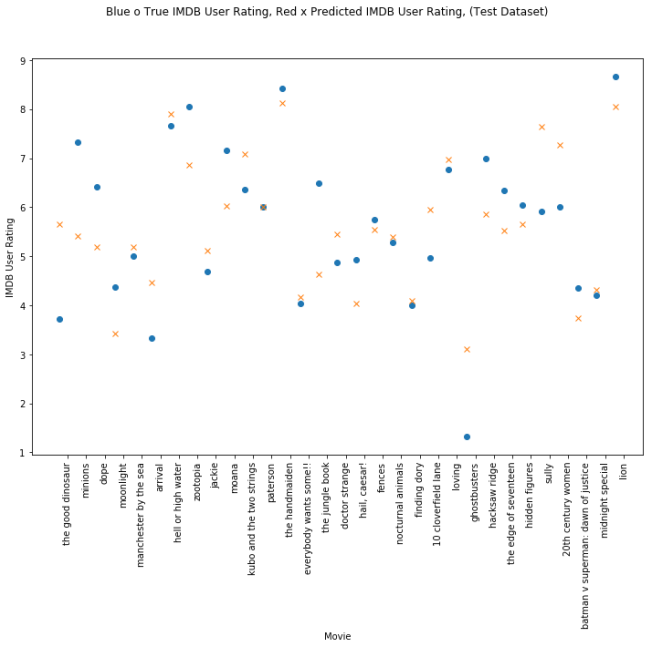

Figure 11: True IMDB User rating vs Predicted IMDB user rating

4 CONCLUSIONS AND FUTURE WORK

We were fairly able to predict the IMDB rating using the developed Multi linear regression model. That shows that social data as a correlation with the success of a movie based on the IMDB score used as a metric.

In future we envision to train the model to predict revenue as well. Currently the data used for training did not involve movies across the world. We plan to make the model robust to predict the success of movies belonging to any movie industry across the world.
We would also like to collect social data for directors and actors for example verified Facebook page likes. The social platform of Twitter needs to be explored as it has abundant user generated content related to movies. Adding such data will only strengthen our given model and out perform our current analysis. Finally having more data, more movies will enhance our model.

ACKNOWLEDGMENTS

This work was supported by Professor Vincent Malic and Ashley Dainas under the course ILS-Z 639 : Social Media Mining. We would like to extend our gratitude towards them for enlightening us with all the concepts of Social Media Mining through a rigorously planned course work.

REFERENCES

[1] Tiberio Uricchio, Francesco Gelli, Marco Bertini, Alberto Del Bimbo and Shih-Fu Chang 2015. Image Popularity Prediction in Social Media Using Sentiment and Context Features *Commun. 23rd ACM international conference on Multimedia* Pages 907-910 http://dl.acm.org/citation.cfm?id=2806361

[2] Sitaram Asur and Bernardo A. Huberman 2010. Predicting the Future with Social Media *Comm. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.

[3] Jure Leskovec 2011. Social Media Analytics: Tracking, Modelling and Predicting the flow of Information through Networks. KDD 2011. http://snap.stanford.edu/proj/socmedia-kdd/

[4] Ming Cheung, James She, Alvin Junus, Lei Can 2016. Prediction of Virality Timing Using Cascades in Social Media. ACM Transactions on Multimedia Computing, Communications, and Applications.

[5] Mestyán M, Yasseri T, Kertész J (2013) Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS ONE 8(8): e71226. doi:10.1371/journal.pone.0071226

[6] Thorsten Hennig-Thurau, Mark B. Houston and Gianfranco Walsh 2006. Determinants of Motion Picture Box Office and Profitability: An Interrelationship Approach. Accepted for publication in Review of Managerial Science December 2006

[7] IMDBPy API: http://imdbpy.sourceforge.net/

[8] IMDB Pie API: https://github.com/richardasaurus/imdb-pie

[9] Facebook API: https://developers.facebook.com/

[10] MovieDB API: https://www.themoviedb.org/documentation/api

[11] Anaconda IDE: https://www.continuum.io/downloads

[12] opencv API: http://opencv.org/

Team Member Participation

	Task	Member(s) Contributed	Comments
1	Google Movies data extraction	Vinita Boolchandani, Madrina Thapa	Complete
2	Rotten Tomatoes Data Extraction	Gurleen Dhody, Vinita Boolchandani	Complete
3	IMDB Data Extraction	Gurleen Dhody, Madrina Thapa	Complete
4	Facebook likes	Gurleen Dhody, Madrina Thapa	Complete
5	Box Office Mojo	Gurleen Dhody, Madrina Thapa	Complete
6	Progress Report Documentation	Vinita Boolchandani, Madrina Thapa	Complete
7	Implementation	Gurleen Dhody, Vinita Boolchandani, Madrina Thapa	Complete
8	Poster Design	Vinita Boolchandani, Madrina Thapa	Complete
9	Project Report	Gurleen Dhody, Vinita Boolchandani, Madrina Thapa	Complete
10	Carrying forward the future work	Gurleen Dhody, Vinita Boolchandani, Madrina Thapa	Complete