A Machine Learning Approach to Predict Movie Box-Office Success

Nahid Quader
and
MD. Osman Gani

Supervisor: Dr. Md. Haider Ali

Co-Supervisor: Dipankar Chaki

Department of Computer Science & Engineering

School of Engineering & Computer Science

BRAC University

Mohakhali, Dhaka-1212

Bangladesh

Thesis Report on

A Machine Learning Approach to Predict Movie Box-Office Success

Author By:

Nahid Quader

ID- 13301028

nahidquader1934@gmail.com

MD. Osman Gani

ID- 13301019

usmansujoy33@gmail.com

Department of Computer Science and Engineering

BRAC University

Supervised By:

Dr. Md. Haider Ali

Professor

Department of Computer Science and Engineering

BRAC University

Co-Supervised By:

Dipankar Chaki

Lecturer

Department of Computer Science and Engineering

BRAC University

Declaration

We declare that, this thesis report is our own work and has not been submitted for any other degree or professional qualifications. All sections of the paper that use quotes or describe an argument or concept developed by another author, have been referenced in the reference section.

…… ……

Nahid Quader MD. Osman Gani

Author Author

BRAC University

Final Reading Approval

Thesis Title: A Machine Learning Approach to Predict Movie Box-Office Success.

The final form of the thesis report is read and approved by Dr. Md. Haider Ali. Its format, citations and bibliographic style are consistent and acceptable. Its illustrative materials including figures and tables are in place. The final manuscript is satisfactory and is ready for submission to the Department of Computer Science & Engineering, School of Engineering and Computer Science, BRAC University.

……..

Dr. Md. Haider Ali
Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Making a prediction of society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis. The motion picture industry is a multi-billion dollar business. And there is a huge amount of data related to movies is available over the internet and that is why it is an interesting topic for data analysis. Machine learning is a novel approach for analyzing data. Our paper proposes a decision support system for movie investment sector using machine learning techniques. In that case, our system will help investors related with this business to avoid investment risks. The system will predict an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomato, Box Office Mojo and Meta Critic. Using different machine learning algorithms, Natural Language Processing and other techniques the system will predict a movie box office profit based on some features like who are the cast and director members, budget, movie release time, various types of movie rating, movie reviews and then process that data for classification.

Key words: Movie Industry, Machine Learning, Support Vector Machine (SVM), Neural Network, Sentiment Analysis.

Acknowledgement

We are grateful to our supervisor, Dr. Md. Haider Ali, for his immense support and valuable ideas throughout the work. He made us to get out of our comfort and push the limit. In every week, he arranged a meeting with us and his precious advises which helped a lot during the research work.

We are also hugely indebted to our co-supervisor Dipankar Chaki. He made a friendly environment which helped us to discuss every problems we faced and overcame during the research work. He was highly active in providing guidelines and his effective concern made us able to finish the research.

Nevertheless, we would like to express our gratitude to department of computer science and engineering, BRAC University and our faculties for helping us with all the necessary support.

Table of Contents

List of Figures

List of Tables

1. Introduction and Overview

Movie industry is a huge sector for investment but larger business sectors have more complexity and it is hard to choose how to invest. Big investments comes with bigger risks. The CEO of Motion Picture Association of America (MPAA) J. Valenti mentioned that 'No one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience' [23]. As movie industry is growing too fast day by day, there are now a huge amount of data available on the internet, which makes it an interesting field for data analysis. Predicting a movie success is a very complex task to do. The definition of a movie success is relative, some movies are called successful based on its worldwide gross income, and some movies may not shines in business part but can be called successful for good critics review and popularity. There are many movies which did not produce good amount of profit during its release time but become famous after few years. For example "Fight Club", a very popular movie of David Fincher released in 1999. Artist like Brad Pitt and Edward Norton were casted for this movie. But this movie was a flop in terms of profit, according to IMDb, budget of "Fight Club" was $63 million but worldwide gross income was only $100 million, which means net profit was only $37 million which is not a good amount of profit at all. But "Fight Club" is a very famous movie now, every movie enthusiasts know the name of this movie, but this movie was not perfect for its time, the movie was a little bit ahead of its time. Now if we consider only profit as a definition of success than "Fight Club" is not a successful movie but if we consider other facts anyone can consider this movie as a successful movie. Again another movie "Transcendence" released in year 2014, budget of this movie was $100 million but total worldwide gross of this movie is $90 million according to IMDb. Star like Johnny Depp was in this movie and this movie got 6.3 rating at IMDb where 187 thousand people voted. This IMDb rating indicates audiences liked this movie but it was a flop. In this paper we considered a movie success based on its profit only. For this type of unpredictable nature of a movie success, it is very confusing decision for

investors to make the right choice. Researches says almost 25% of movie revenue comes within the first or second week of its release [24]. So it is hard to predict a move success before its release.

Our mission of this paper is to make a model which can help investors to avoid risks and make a right choice of investment. This research will not only help investors but also will be helpful for the whole movie industry. There are many new artist who cannot make a film because no investor is ready to invest for them. Investors has their own reason, not all investor has the courage to invest on a movie of a new director because he/she has no experience to show but they are extremely talented and passionate about film making. Early prediction will help an investor to make choice if he/she wants to invest for new artists. This will be great for new artist in the movie industry. A movie industry contributes a massive amount of money in global economy, everything is connected now in 2017. So if new artists can make movie easily more artist will try to make films, more films will produced day by day and movie industry will contribute more money to global economy. Our mission is to help investors to easily make choices but the vision or the main goal is to help the movie industry of Bangladesh in the future. In Bangladesh there are many new artist who are willing to make new films but cannot do anything because of money, no producers are ready to give money to those new artists. For film industry of Bangladesh this research will be very helpful because in our country producers does not have enough money to make a blind bet on some new artists, so if we can give them an idea about how well a movie can do business after release it will be extremely helpful for them. We have made our dataset based on foreign movies only because unavailability of necessary information on the internet for movies in our country. To manage all data of movies released in our country will take so much time, this is our main goal in future to collect data and make prediction to help the film industry of Bangladesh.

In our proposed model we have used two types of features called pre-release features and post-release features. To predict an upcoming movie only pre-release features will be

responsible for prediction. To predict right after release, both pre-release and post-release features will be responsible. There are six pre-release features and nine pre-release features. More features helped us to make good and more generalized prediction. Instead of forecasting only flop or blockbuster movies [10], we rather choose to classify a movie based on its box office profit in one of five categories ranging from flop to blockbuster (Table 4.2). For multiclass prediction several machine learning algorithms are available like Naive Bayes, Support Vector Machine (SVM) and Logistic Regression etc. These classifiers are good enough for binary classification, some of them can be used for multi class classification but when data pattern is very complex, Neural Network consistently produce better result. We applied both SVM and Neural Networks on our dataset for prediction, among these two methods Neural Network produced comparatively good result.

The reminder of this paper is organized as below:

- In section 2, Literature Review of previous research works.
- In section 3, Methodology, system work flow, Support Vector Machine (SVM) and Neural Network definition.
- In section 4, Data Description, Data acquisition, Data cleaning, Feature extraction, Data integration and transformation.
- In section 5, Analysis and Result discussion, sentiment analysis, SVM and Neural Networks performance analysis.
- In section 6, Conclusion and Discussion.
- In section 7, References

2. Literature Reviews

Success of a movie primarily depends on the perspectives how the movie has been justified. In early days, a number of people prioritized gross box office revenue ([1], [2], [3], [4]), initially. Few previous work ([4], [5], [6]), portend gross of a movie depending on stochastic and regression models by using IMDb data. Some of them categorized either success or flop based on their revenues and apply binary classifications for forecast. The measurement of success of a movie does not solely depend on revenue. Success of movies rely on a numerous issues like actors/actresses, director, time of release, background story etc. Further few people had made a prediction model with some pre-released data which were used as their features [7]. In most of the case, people considered a very few features. As a result, their models work poorly. However, they ignored participation of audiences on whom success of a movie mostly depends. Although few people adopt many applications of NLP for sentiment analysis ([8], [9]) and gathered movie reviews for their test domain. But the accuracy of prediction lies on how big the test domain is. A small domain is not a good idea for measurement. Again most of them did not take critics reviews in account. Besides, users' reviews can be biased as a fan of actor/actress may fail to give unbiased opinion.

M. T. Lash and K. Zhao's [10] main contribution was, firstly they developed a decision support system using machine learning, text mining and social network analysis to predict movie profitability not revenue. Their research features several features such as dynamic network features, plot topic distributions means the match between "what" and "who" and the match between "what" and "when" and the use of profit based star power measures. They analyzed movie success in three categories, audience based, released based and movie based. Their hypothesis based on the more optimistic, positive, or excited the audiences are about a movie, the more likely it is to have a higher revenue. Similarly, a movie with more pessimistic and negative receptions from the public may attract fewer people to fill seats. They retrieve data from different types of media. Such as Twitter, comments from YouTube, blogs, new

articles and movie reviews, star rating from reviews, the sentiment of reviews or comments have been used as a means for assessing audience's excitement towards a movie. Their original dataset collected from both BoxOfficeMojo and IMDb. They focused on the movies released in USA and excluded all foreign movies from their experiment.

In [11] A. Sivasantoshreddy, P. Kasat, and A. Jain tried to predict a movie box-office opening prediction using hype analysis. Mainly this paper is focusing on twitter data for hype analysis. Main logic behind hype analysis is a success of a movie heavily depend on its opening weekend income and also how much hype it gets among people before release. At first they found the number of tweets pertaining to a movie by using web crawler. These tweets are collected by hour basis. There are three factors for hype measurement. First factor is to calculate "No of relevant tweets per second." Second factor is "Find the number of distinct users who have posted the tweets". Third factor is "Calculate the reach of a tweet". Here reach of a tweet means that some different person's tweets have different value. Suppose if a well-known actor or director posted a positive tweet for a movie is more valuable than a tweet posted by an average person. For calculating the reach of a tweet they count the follower of a particular user. They calculated No of relevant tweets per second, Second factor is "Find the and Calculate the reach of a tweet as hype factor by taking the average value of these three factors for each movie. Their analysis based on hype factor, number of screens the movie is going to be released and the average price of all tickets per screen per show. The total model is very simple calculations and they just counted the number of tweets related to a movie, but they don't use any kind of language processing to know if the tweet is positive or negative. A neural network had been used in the prediction of financial success of a box office movie before releasing the movie in theaters [12]. This forecasting had been converted into a classification problem categorized in 9 classes. The model was represented with very few features.

In [13], it was tried to improve movie gross prediction through News analysis where quantitative news data generated by *Lydia* (high-speed text processing system for collecting and analyzing news data). It contained two different models (regression and *k*-nearest neighbor models). But they considered only high budget movies. The model failed if common word used as name and it could not predict if there were no news about a movie. M.H Latif, H. Afzal [14] who used IMDB database only as their main source and their data was not clean. Again their data was inconsistent and very noisy as they mentioned. So they used Central Tendency as a standard for filling missing values for different attributes. K. Jonas, N. Stefan, S. Daniel, F. Kai use sentiment and social network analysis for prediction [15] their hypothesis was based on intensity and positivity analysis of IMDb's sub forum Oscar Buzz. They had considered movie critics as the influencer and their predictive perspective. They used bag of word which gave wrong result when some words were used for negative means. There was no category award and only concerned with the award for best movie, director, actors/actress and supporting actors/actress. In some cases, success prediction of a movie were made through neural network analysis ([7], [18]). Some researchers made prediction based on social media, social network and hype analysis ([16], [17], [19], [20]) where they calculated positivity and number of comments related to a particular movie. Moreover few people had predicted Box Office movies' success based on Twitter tweets and YouTube comments. In both case, the accuracy of prediction will be doubtful and will fail to give appropriate result. A small domain is not a good idea for measurement. In previous works, most researches were based on attributes that were either available prior to the release or after the release of a movie. Although some of the researchers had considered both types of attributes but in that case very few attributes were counted. The possibility of having better success in prediction goes higher with more attribute involved.

3. Methodology

3. 1. Workflow

The first phase is data acquisition. Here we choose four data sources IMDb, Rotten Tomato, Box Office Mojo and Meta Critic. Different types of features are extracted from different sources which will be described thoroughly in section 5. Second phase is data cleaning. After scrapping data from various sources, we cleaned our data mainly depend on unavailability of some features. After cleaning all data, next phase is data integration and transformation. In third phase we classified some features. Specific classification details are shown in section 5. Fourth phase is Sentiment analysis of IMDb and Rotten Tomato reviews. Microsoft text analytics API and Power Bi tool has been used for sentiment analysis. Sentiment analysis value has been integrated in dataset with multiplied by the number of reviews. Fifth phase is Result and Analysis, where we applied Support Vector Machine (SVM) and Neural Network on our dataset.

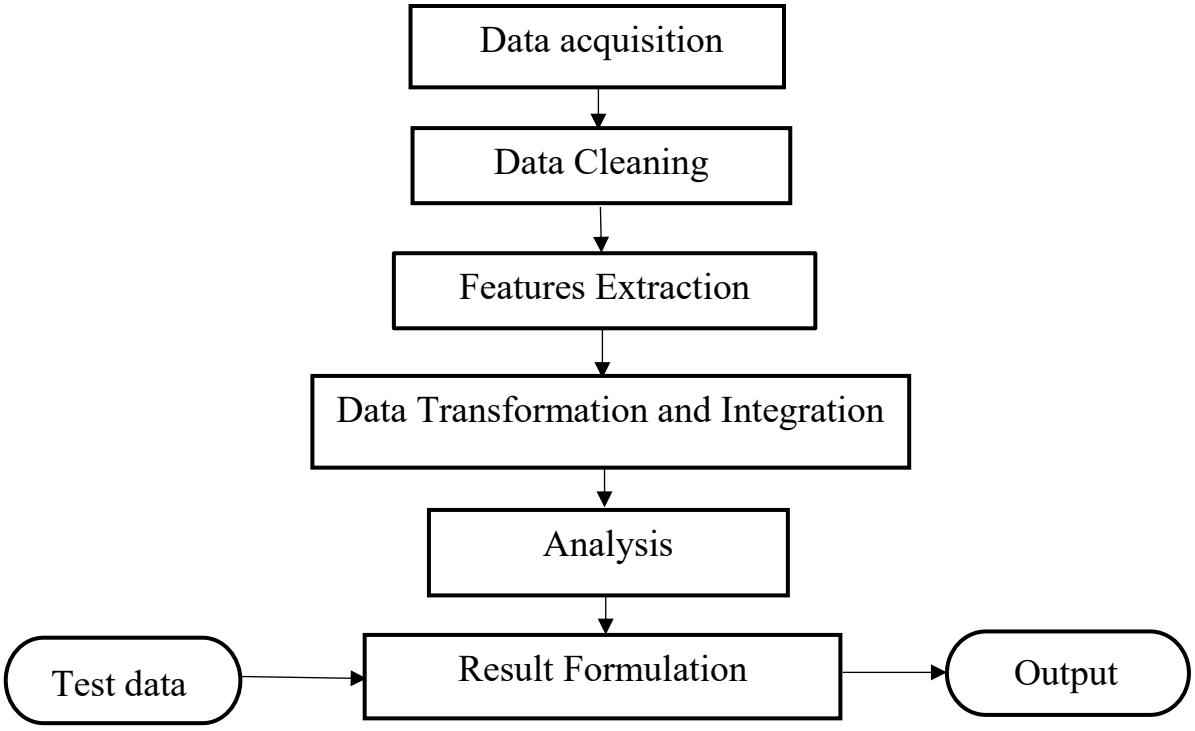


Figure 3.1: Research Workflow

The workflow of the research is given as flowchart in Fig. 3.1.

3. 2. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a classifier formally defined by a separating hyperplane. The goal is to design hyperplanes that classifies all training vectors in classes. Here in this example we have two different features and two classes. We show two different hyperplanes which can classify correctly all the instances in this features have but the best choice will be the hyperplane that leaves the maximum margin from both classes. The margin is the closest distance of elements from the hyperplane. In Fig 3.2, for the red hyperplane Z1 is the margin and Z2 is the margin of green hyperplane. We can clearly see that the margin value Z2 is higher than Z1.

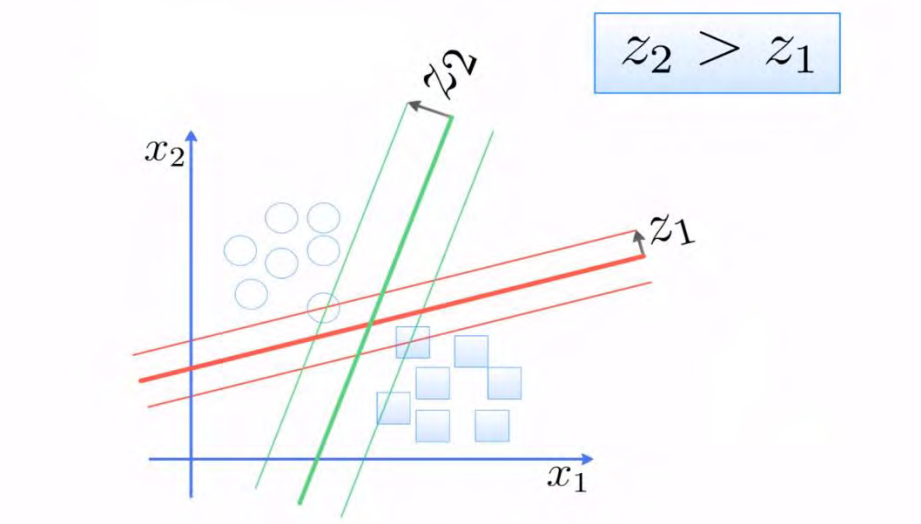


Figure 3.2: SVM Hyperplanes

So the margin of green hyperplane is high and the best choice will be the green hyperplane. This green hyperplane is defined by the equation (1) where $\vec{\omega}^T$ is a vector of weights.

$$g(\vec{x}) = \vec{\omega}^T.\vec{x} + \omega_0 \quad \begin{matrix} g(\vec{x}) \geq 1, & \forall \vec{x}\ \varepsilon\ \text{class 1} \\ g(\vec{x}) \leq -1, & \forall \vec{x}\ \varepsilon\ \text{class 2} \end{matrix} \tag{1}$$

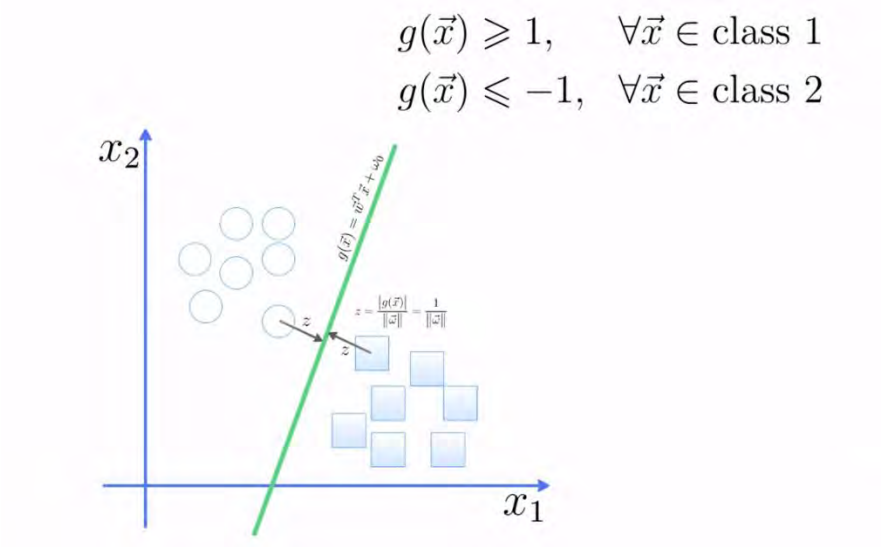


Figure 3.3: Distance of hyperplane to closest elements

This equation will deliver values greater than 1 for all the input vectors which belongs to the class 1 which is circles and also values smaller than -1 for all values which belongs to the class 2 which is rectangles. From Fig. 3.3, we can say that the distance from hyperplane to the closest elements will be at least 1. And from the geometry we know that the distance between a point and a hyperplane is computed by the equation (2). So the total margin which composed by the distance will be computed by the equation (3).

$$Z = \frac{|g(\vec{x})|}{\|\vec{\omega}\|} = \frac{1}{\|\vec{\omega}\|} \tag{2}$$

$$\frac{1}{\|\vec{\omega}\|} + \frac{1}{\|\vec{\omega}\|} = \frac{2}{\|\vec{\omega}\|} \tag{3}$$

And the aim is to that minimizing the term $\|\vec{\omega}\|$ in $\frac{2}{\|\vec{\omega}\|}$ in the equation (3) which will maximize the separability. When we minimize the weight vector $\|\vec{\omega}\|$ we will have the biggest margin z here that will split all the classes. This classification method of SVM is also called Support Vector Classification (SVC). For implementation of SVM, we have used python as our programming language and its' machine learning library Scikit-Learn [21] which designed to

interoperate with Numpy and SciPy. NumPy and SciPy are pythons' scientific and numerical libraries. Scikit-learn is being currently funded by INRIA (Inventors for the Digital World), Paris-Saclay Center for Data Science, NYU Moore-Sloan Data Science Environment, Telecom Paristech, Columbia University, Alfred P. Sloan Foundation and The University of Sydney.

For visualization we have used pythons' Matplotlib library. In our analysis, we apply SVC with kernel linear, kernel Gaussian Radial Basis Function (RBF), kernel polynomial and LinearSVC. The main difference between SVC with kernel linear and LinearSVC is their implementations. LinearSVC is based on liblinear and SVC kernel linear is based on libsvm library. LinearSVC is more flexible to choose the penalties and loss functions and better on large numbers of samples. Moreover, SVC kernel linear uses one vs one scheme when LinearSVC uses one vs rest scheme for classification. By selecting best parameters we have found different accuracy in different method. We applied SVM on only pre-released features and post-released features with pre-released features.

3. 3. Neural Network

Prediction accuracy of Neural Networks on our dataset is best among all other classifiers. Neural Network is like any other kind network, there are interconnected web of nodes which are called neurons and edges which join them together. Neural nets receives a set of inputs, perform progressively complex calculations and gives output based on its calculations. There are different types of classifiers available like Logistic Regression, Support Vector Machine (SVM) and Naive Bayes. Neural networks are best for pattern recognition. For analyzing simple patterns, the basic classifiers like SVM or Naive Bayes works great, but for more complex pattern with more than ten inputs, Neural Networks start to perform better than other methods. Neural Networks are highly structured and comes in layers. For more complex patterns a Neural Network need more layers because the number of nodes require for each

layer grows exponentially with the number of possible patterns in data. The first layer called input layer, the final layer called output layer and layers between them are hidden layers. Each node has its own activation function. One node takes input from previous nodes and activates and the score is passed on as inputs to next layer for further activation until it reached the output layer. This series of events starting from the input and send to the next layers for activations all the way to the output is called forward propagation. Every nodes in a Neural Network is connected with each other. Each node has its own activation functions and none of them gives random output. That means if the same input is given repeatedly it will give same output. Doing this will definitely not help for pattern recognition. But the actual scenario is not like this, Neural Network gives different value for each epoch. The reason is each set of inputs are unique by its own weight and biases. This means the combination used for each activation is unique. The main goal is to change these weights to achieve least cost or more accuracy. The process of improving a Neural Network's accuracy is called training. To train the net, the output from forward propagation is compared with the actual output in given data. The cost is simply the difference of generated output and actual output equation (4).

$$\text{Cost} = \text{Generated Output} - \text{Actual Output} \tag{4}$$

There is where back propagation comes. For each train, weights and biases are changed to improve accuracy and this changes is done by back propagation. Changing weights are very important for training because the main idea of training or learning is to adjust the weights to make the error as low as possible. So if the weights are changed randomly it will take so much time to reach the goal or maybe it will never reach, so it is not practical. To figure out the direction (higher or lower) that we need to adjust the weight by calculating the slope. If we plot the error and weights in a graph (Fig. 3.4) we will see that there is a sweet value of weight where the error is minimum. To know the direction we have to calculate the slope which is shown in equation (5).

Slope=Change in weights / Change in Error (5)

By calculating this slope weight can be adjust throughout the training.

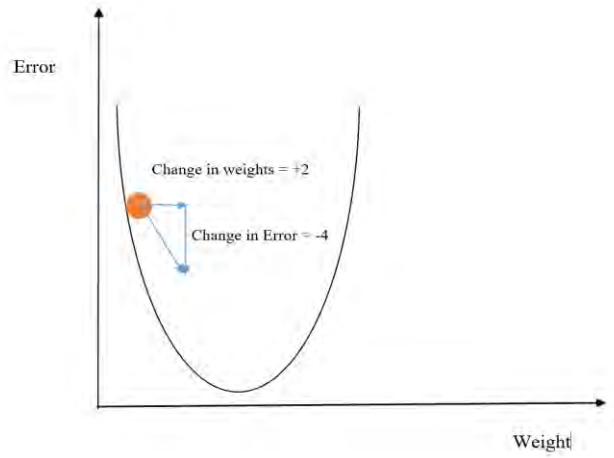


Figure 3.4: Calculating the direction of changing weights

Now the activation functions are very important. There are different types of activation functions like Sigmoid, ReLu, Softmax etc.Different Activation functions gives outputs in different range.

The range of sigmoid is 0 to 1 (Fig. 3.5). It maps every input in between this range and gives output. But the problem of this function is, it is not zero centered and vanishing gradient. So for weight update, it sometimes goes too far in different direction. Sigmoid works god for binary classifications.

Rectified Linear Units (ReLu) is a very famous activation function (Fig.3.6). The function is, ReLu(input) = max(0,input) if input < 0 ,ReLu = 0 or ReLu = input.Its simple and efficient but for only hidden layers.

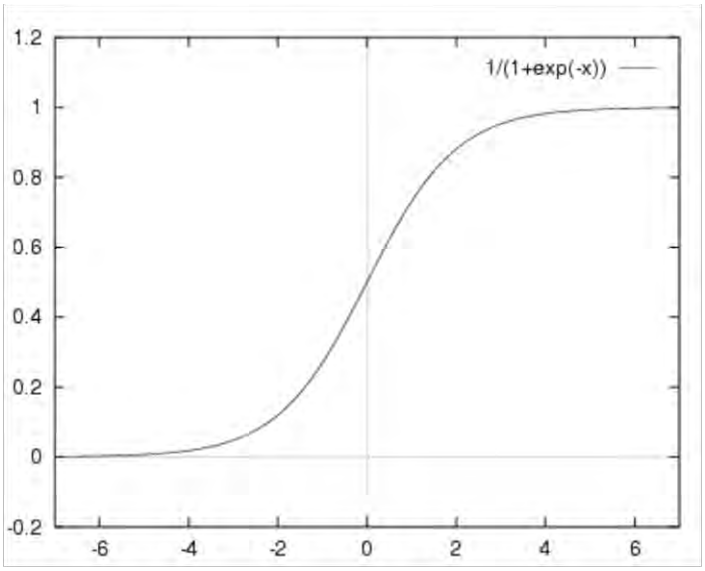


Figure 3.5: Sigmoid Activation Function

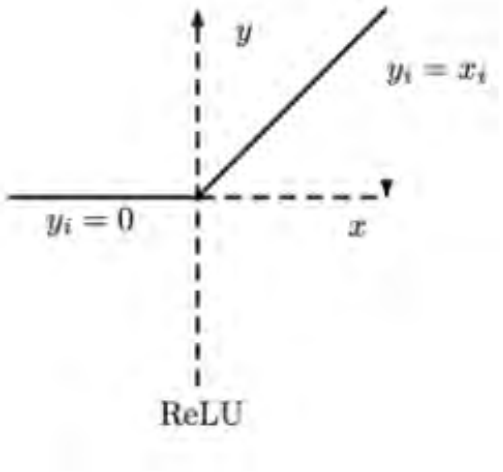


Figure 3.6: ReLu Activation Function

Softmax is a very good choice for multi class prediction problem. Softmax gives a good result in case of compute probabilities for classes. The highest probability of a class will be the final prediction class.

4. Data Description

4. 1. Data Acquisition:

This dataset contains 755 movies released in between 2012 to 2015 (Fig. 4.1). Recent movies are not selected because movie information are changing every day. Main data sources are IMDb, Rotten Tomato, Metacritic and Box Office Mojo. We took IMDb rating, MPAA, IMDb Votes, Genre, Directors, Casts and Country using IMDbPy library available in python. But IMDbPy had some issues, they don't provide business data.

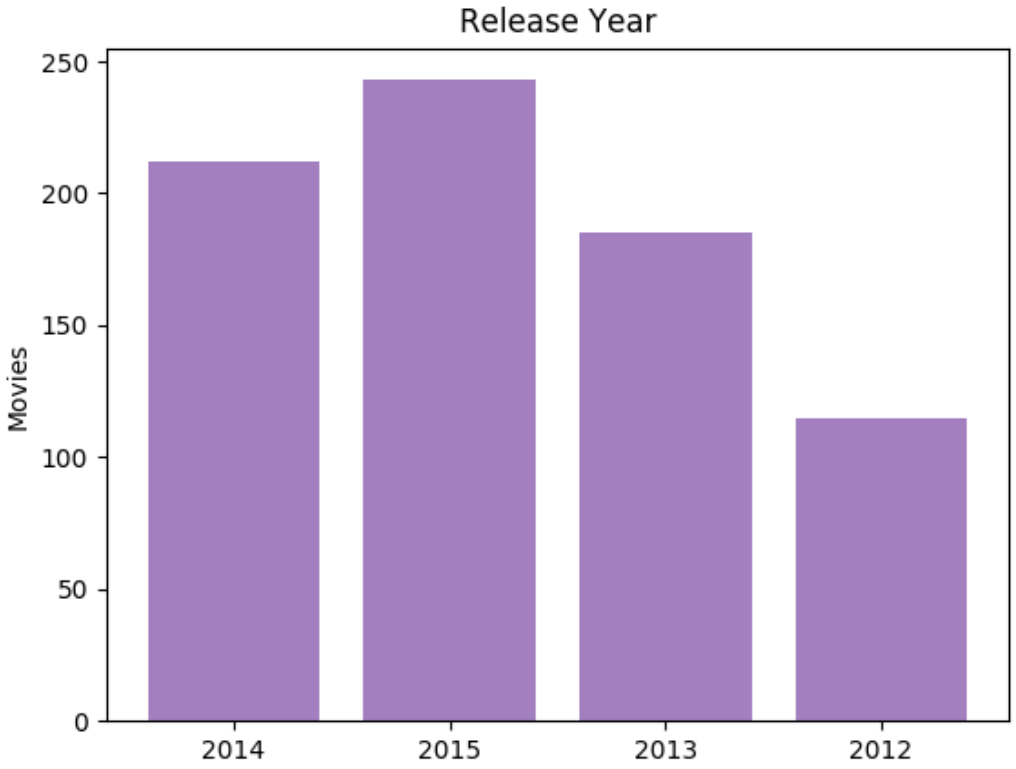


Figure 4.1: Number of movies in each year

So we couldn't get budget and movie gross income from IMDbPy. We had to scrap IMDb website to get those features. IMDb has a very well managed website. They use a unique

id for both movies and artists. These unique id was pulled from imdbpy library. For star power calculation, star gross income were extracted from IMDb website. Metacritic rating is a very important feature which have been taken from Metacritic website. Two types of reviews are used in this dataset. One is audience reviews which has been taken from IMDb audience review section and critic reviews from Rotten Tomato critic review section. From Rotten Tomato we took Tomato meter, Tomato rating which are given by movie critics and Audience meter, Audience Rating are given by audiences. We faced severe problem while scraping Rotten Tomato website. Unlike IMDb website Rotten Tomato uses movie or artist's names to create their links. It was very hard to open links autonomously because there was no pattern like IMDb. For some movie links they used movie release year, for some movies the year was not correct. Even if in some link they used some arbitrary codes which doesn't make any sense. So scraping in Rotten Tomato was a big deal. But we have done it successfully. Move budget value are taken from Box Office Mojo.

Table 4.1: Dataset Summary

Features	Type	Mean	Median	Min	Max	Std. Dev	Data Source
IMDb Rating	Float	6.44	6.5	1.5	8.6	0.91828	IMDb
Tomato Meter	Float	51.9735	52	0	99	28.6999	Rotten Tomato
Tomato Rating	Float	5.5404	5.7	0	9.2	1.79299	Rotten Tomato
Audience Meter	Float	57.8558	58	0	94	19.0728	Rotten Tomato
Audience Rating	Float	3.4108	3.45	0	4.5	0.53570	Rotten Tomato
Meta Score	Integer	50	52	0	100	20	Meta Critic
MPAA	Integer	3.1284	3	0	4	1.0952	IMDb
Cast Star Power	Integer	9606441175	7999583961	0	50943162024	7846413921	IMDb
IMDb Review Sentiment (Multiplied by no of reviews)	Integer	172	91.71049	0	2298	246	IMDb
Rotten Tomato Review Sentiment (Multiplied by no of reviews)	Integer	75	67.9017	0	264	54	Rotten Tomato
IMDb Votes	Integer	110633	55918.5	655	1201640	148731	IMDb
Release Month	Integer	6.6251	7	1	12	3.4222	IMDb
Budget	Integer	41492981	20000000	20000	250000000	51823867	BoxOfficeMojo, IMDb
Number of Screen	Integer	1884	2275.5	1	4404	1549	Box Office Mojo
Director Star Power	Integer	1132921173	417051548	1441	13913175395	1741929261	IMDb

4. 2. Data Cleaning:

At first our dataset had 2761 movies. Then we recognize that there were many movies which doesn't have all data available. So unavailability of features was the main reason behind eliminating movies from our dataset. Most of the movie doesn't have budget data available.

We checked IMDb first. When we saw IMDb doesn't have budget for a movie, we searched in Rotten Tomato, Box-Office Mojo for budget. For some movie we got budget from Box-Office Mojo but for most of the movies in our dataset, budget was unavailable. After removing those movies there was 800 movies left. Some movies which had all necessary business information like budget and gross income but no other information was available, For some movies we couldn't calculate star power because there was no data available about those casts on the internet. Most of these movies were Indian. After removing those movies we finally got our dataset with 755 movies which has all information available. Table 4.1 shows the summary of our dataset.

4. 3. Features Extraction

In previous works, very few features were considered in most of the models. This paper included most of the features available in online. Two types of features are considered in this paper, one is pre-released features and post-released features. Only the pre-released features are available for upcoming movies to predict the success. Here pre-released features are the budget of a movie, the number of screens where the movie will be released, Motion Picture Association of America (MPAA) rating, actors/actress's star power, start power of director and the month of the release. Furthermore, after one or few weeks of releasing a movie, post-released features will be useful to make a better accuracy in prediction as those will be available. IMDb rating, Rotten Tomato Critics' Meter and Rating, Rotten Tomato Audiences'

Meter and Rating, Meta Score, sentiment of IMDb and Rotten Tomato reviews with the number of reviews, IMDb votes.

4. 3. 1. Rating and Votes:

It took in consideration of tomato critics' meter, tomato critics' rating, tomato audiences' score and tomato audiences' rating from Rotten Tomatoes, Meta score of Metacritic and IMDb rating from IMDb. Each movie in Rotten Tomatoes has tomato critics' meter which is evaluated by movie critics given in percentage and tomato critics' rating which is also given by movie critics from 0 to 10. Again it has tomato audiences' percentage score and tomato audiences rating from 0 to 5 given by movie audiences. Like Rotten Tomatoes, IMDb also a rating from 0 to 10. Further this paper also counted how many audiences voted on IMDb to make the rating. IMDb shows average of all the votes as their rating for a particular movie. The higher number of votes indicates more people have watched that particular movie. Every user can vote any movie and average of the audiences' voting shown as users' rating in their official website. Another rating is Meta score which is also evaluated by only movie critics published in Metacritic official website ranging from 0 to 100.

4. 3. 2. MPAA:

Motion Picture Association of America (MPAA) is a governing body which rate a movie's suitability for certain audience based on its content. This system is a voluntary scheme and many theaters refuse to exhibit non-rated movies. In Fig. 4.2, there are total 5 categories for each of a movies which are R, PG, PG13, G and NC. As it is a voluntary scheme and it is not enforced by law that's why there are some movies without MPAA rating. This paper rated them as NC which meant not rated by MPAA.

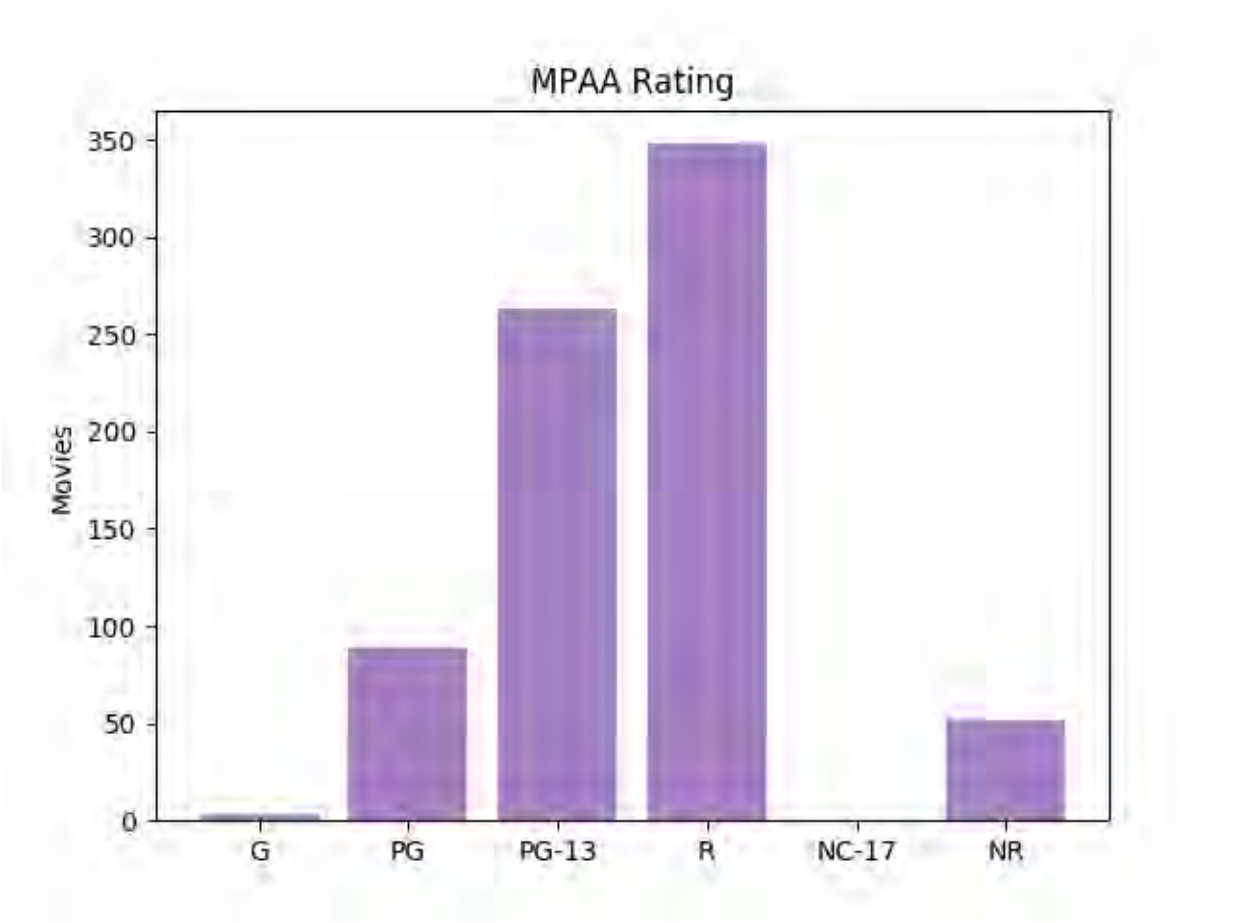


Figure 4.2: MPAA Rating

4. 3. 3. Star Power:

In next, this paper considered the star power of actors, actresses and directors. Celebrity actors and directors like Brad Pitt, Tom Hanks, George Clooney and Quentin Tarantino are well known throughout the movie audiences. Famous artists like them not only make high quality movies but also increase the success probability of their movies and there will be a very large number of people who will be keenly interested to watch that movie. Cristopher Nolan, the director of Memento (2000), the prestige (2006), The Dark Knight (2008), Inception (2010) and Interstellar (2014), can be considered as a star director as his most of the movies earned well reputations and very high worldwide gross. So it can be said, the next movie of Cristopher Nolan will be profitable. Again Leonardo DiCaprio is very famous for his movies,

Titanic (1997), The Revenant (2015), The Wolf of Wall Street (2013), Inception (2010) and many more. It will be wise to say his next movie will get very high income and popularity. This paper considered an actor/actress as a star whose previous movies got high income. If an actor/actress' previous movies have high box office gross, it can be said he/she became familiar among audiences. And then by calculating total gross of all the movies from his/her whole career, it can be used as a parameter for his/her popularity. This paper considered the total gross of all the movies of an actor/actress/director in their career as their star power. The more popular, an actor/actress/director is, the more successful movies he/she has.

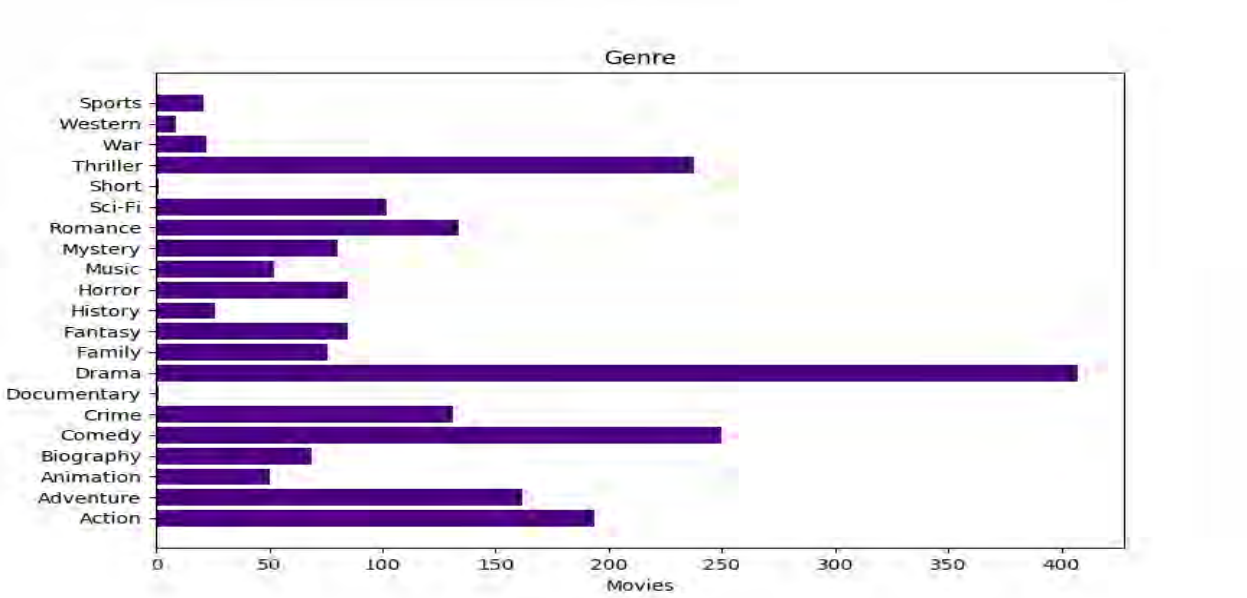


Figure 4.3: Genre

4. 3. 4. Month of Release:

In addition, release date is also a big factor in the business of motion picture industry. It is seemed to have more crowds in theaters for a movie then it is being released in public holyday. In this article, release date is considered as month of movie release. Fig. 4.4 shows the number of movies according to the month of released.

4. 3. 5. Budget:

Budget is another pre released data which has very important effect on the prediction of movie box office success. If a movie has higher budget for making the film and for its publicity, it has high chance to get more hypes. So higher budget movie has higher chances to income more. Budget has been calculated with the inflation adjustment. For instance, Superman (1978) had a budget of 55 million USD without inflation adjustment but after inflation adjustment it had more than 200 million USD. We have collected budget of movies from IMDb and Box Office Mojo and adjusted inflation.

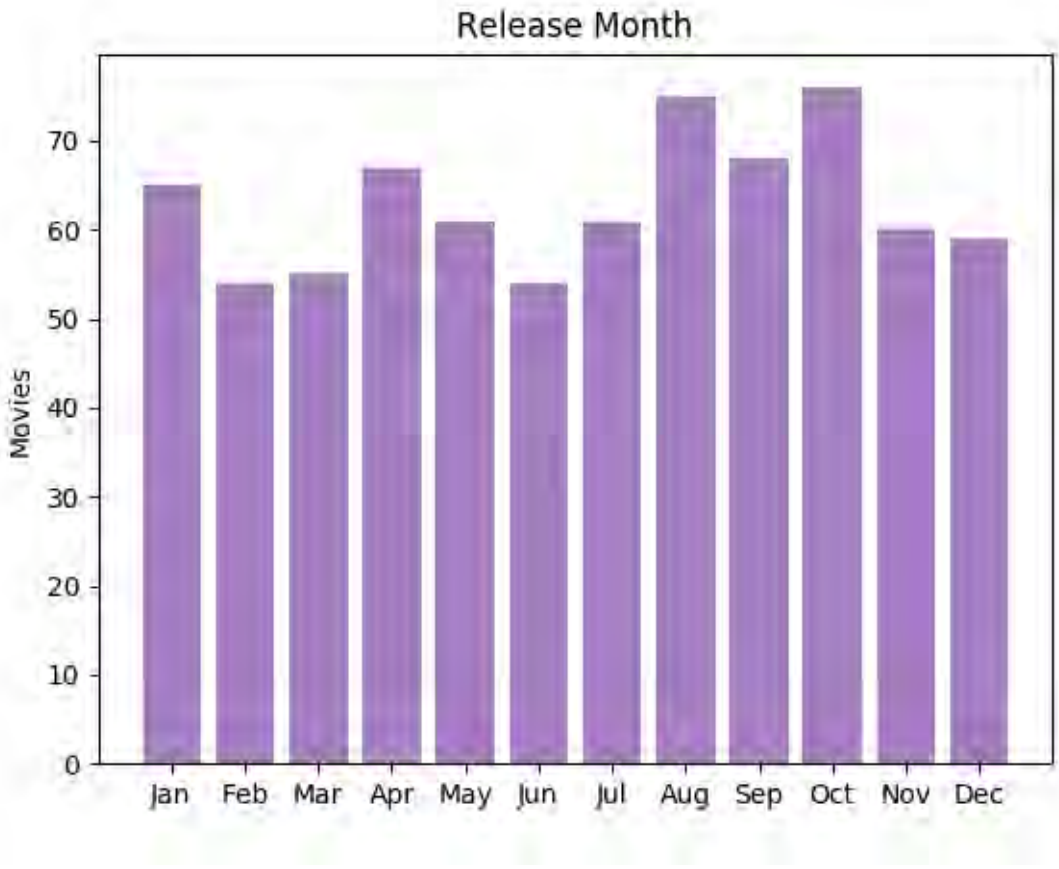


Figure 4.4: Month of Release

4. 3. 6. No. of Screens:

If a movie is released in more number of screens, the more numbers of people will be able to watch the movie and the more number of tickets will be sold. Thus number of screens

has a great effect in the financial success of a movie in the box office. This feature has been added from Box Office Mojo.

4. 3. 7. Reviews:

Further we had taken critics reviews into account from Rotten Tomato along with the IMDb's and Rotten Tomato's users' reviews. We prioritized critics' reviews as critics focus on movie's story and over all performances of the cast members. In addition, users' reviews can be biased as a user can be fan of any actor/actress/director. On the other hand critics show professional judgement in their reviews. Hence prioritizing critics' review rather than users' reviews will help to get more accurate result in prediction. In the historical analysis, IMDb vote, meta score of Metacritic and tomato meter, tomato rating and tomato users' of Rotten Tomato will help us in having more accurate forecast of a movie's success. To understand which movie has created more excitement among audiences the number of reviews for both IMDb and Rotten Tomato has been listed.

Table 4.2: Target Class

Target Class	Range (USD)
1	Profit <= 0.5M (Flop)
2	0.5M < Profit <= 1M
3	1M < Profit <= 40M
4	40M < Profit <= 150M
5	Profit > 150M (Blockbuster)

4. 4. Data Integration and Transformation:

Profit is the most important thing in this model. Here profit is the main target class. In historical analysis budget and gross of movies are collected from IMDb and BoxOfficeMojo. Profit is calculated by subtracting budget from total gross. In addition, profits were calculated with Inflation adjustment. Inflation is an important factor as value of money is changing every day. For instance, Avatar (2009) is the highest-grossing film of all-time with $2,787 million but after inflation adjustment Gone with the wind (1939) becomes the highest-grossing with $3,440 million worldwide gross and Avatar (2009) is in the second position with $3,020 million worldwide gross. After cleaning, all features have been classified. There are five target classes based on the amount of profit a movie made. If profit is less than half million USD, that movie is considered as flop. On the other hand a blockbuster movie has profit more than one fifty million USD and classified as class 5. Table 4.2 shows the target class classification. IMDb (Table 4.3) rating and Rotten Tomato Critics ratings are classified by five classes. Class one means poor rating. There are very few successful movies which got four or less IMDb rating. A movie with more than 7.5 IMDb rating is excellent and got class value four.

Table 4.3: Classification of IMDb Rating

Class	IMDb Rating (Out of 10)
1	Rating <= 5
2	5 < Rating <= 6.5
3	6.5 < Rating <= 7.5
4	7.5 < Rating

Rotten Tomato Critic Meter and Audience Meter are classified in 3 classes. In Rotten Tomato website they mentioned that, to achieve an overall good review score 'Fresh Tomato', a movie has to get at least 60%. To get a 'Certified Fresh Rating' a movie need 75% or higher. If score is less than 60% it is called 'Rotten Tomato'. Table 4.4 is given by Rotten Tomato.

Table 4.4: Tomato Meter Classification Defined by Rotten Tomato

Overall rating	Meter Score(in percentage)
Rotten Tomato	Score < 60
Fresh Tomato	60 < Score < 75
Certified Fresh Rating	75 <= Score

Rotten Tomato also mentioned about score, above 3.5 rating means 'The Full Popcorn Bucket' unless 'The Tipped Over Popcorn'.

Table 4.5: Meta Score Classification

Class	Meter Score(in percentage)
1	Score < 60
2	60 < Score < 75
3	75 <= Score

Meta score has been taken from Metacritic where only professional critics gives their opinion and rating. In table 4.5, Meta score has been classified. Audience score and critics score is classified by following Rotten Tomato's classification (Table 4.4) and their rating classification is given in table 4.6. MPAA (Motion Picture Association of America) has six ratings G, PG, PG-13, R, NC-17 and NR (Not Rated. MPAA has been classified in 6 classes (Table 4.7).

Table 4.6: Audience and Critics Rating

Class	Score (Out of 5)
1	Score < 3.5
2	Score => 3.5

Release Date is classified my months. We wanted to see if there any relation between movie income and release month. For twelve months there are twelve classes. Star power has been calculated in two way, one is actor gross value another one is director gross value. Both actor and director gross value is calculated in same way. Actors/Actress star power (Table 4.8) is classified in five classes, the classes are defined by the star value. Similarly director is also important as much as actors/actress power. So we have classified directors' star power in the same way actors/actress star power have been classified. And Director star power is classified by 3 classes in Table 4.9.

Table 4.7: MPAA Rating

Class	MPAA Rating
0	NR
1	G
2	PG
3	PG-13
4	R
5	NC-17

Table 4.8: Actors/Actress Star Power Classification

Class	Range (in billions)
1	Star power < 3
2	3 <= Star power < 7
3	7 <= Star power < 10
4	10 <= Star power < 15
5	Star power => 15

Table 4.9: Director Star Power Classification

Class	Range (in millions)
1	Star power < 100
2	100 <= Star power < 1000
3	1000 <= Star power < 10

Number of screen is a very important feature of our dataset. A movie business depends on the number of screen released. The widest release screen numbers are taken from Box Office Mojo. Number of screens is classified in five classes (Table 4.10).

Table 4.10: No. of Screen Classification

Class	Range
1	No of Screen <= 100
2	100 < No of Screen <= 500
3	500 < No of Screen <= 2000
4	2000 < No of Screen <= 3000
5	No of Screen > 3000

5. Analysis and Result

5. 1. Sentiment Analysis:

In total 212535 reviews from IMDb and 108464 reviews from Rotten Tomato are collected in the dataset. Sentiment values of those are calculated with Microsoft Power Bi Desktop application. With Power Bi Desktop we used Microsoft Azure's cognitive service of Text Analytics API. Text analytics API is a Natural Language Processing application which calculates the positivity and negativity of a text. They provide 5000 free transitions of grouped data for their services and every group consists of maximum 1000 individual data. It gives sentiment value ranging from 0 to 1 where sentiment value of a review close to 1 means highly positive review, 0 means highly negative and 0.5 means neutral review. We made 213 transition for IMDb and 109 transition for Rotten Tomato manually through Power Bi Desktop. After getting those sentiment value for each reviews, we calculated the mean value of reviews for every movie individually along with the number of reviews for each movie. A positive sentiment mean value does not the movie is successful. Sometimes an unsuccessful movie can have very few reviews with highly positive sentiment mean value, on the other hand a highly successful movie can have lots of views with poor sentiment mean value. For instance, Interstellar (2014) has 0.7064 mean sentiment value with 2830 reviews and The Theory of Everything (2014) has 0.8760 mean sentiment value with 407 reviews in IMDb but Interstellar (2014) has been classified in profit class 5 and The Theory of Everything (2014) in 4. For this reason the number of reviews for a particular movie has been listed as it will help us to understand which movie has created more excitement among the audiences. In next we have multiplied the review counting number with the sentiment mean value and listed the result as review sentiment value in our dataset for IMDb and Rotten Tomato and used them as our features.

5. 2. Support Vector Machine (SVM):

We have implemented 10 fold cross validation in each of our experiments. In 10 fold Cross validation, all the elements in our dataset are divided into 10 groups.

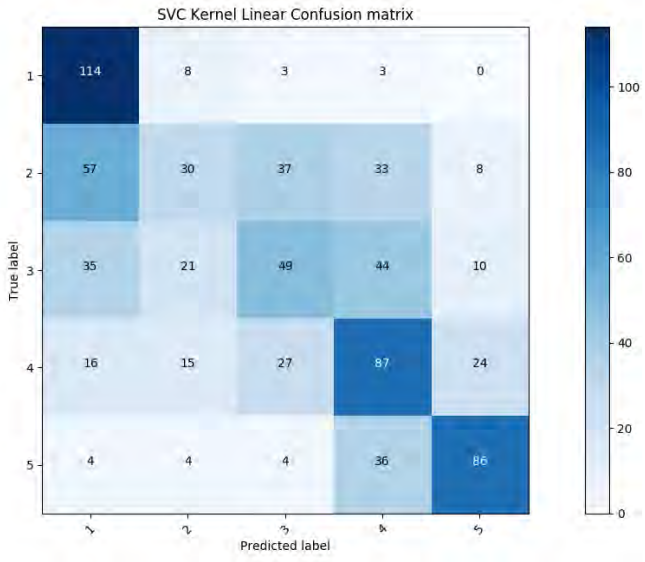


Figure 5.1: SVC Kernel Linear (Pre-Released Features)

Table 5.1: SVC Kernel Linear (Pre-Released Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	45.45	53.24	49.35	57.14	44.73	46.66	49.33	48.00	41.89	48.61	48.44

In next first group becomes the test data and rest nine groups make the train data for the machine and we listed its accuracy. After testing first group, second group becomes the testing data and rest groups make the training data for the machine. In this way all data are

tested and mean is calculated from the accuracy of each fold. In SVC with kernel linear on pre-released features we have mean accuracy of 48.44 percentage.

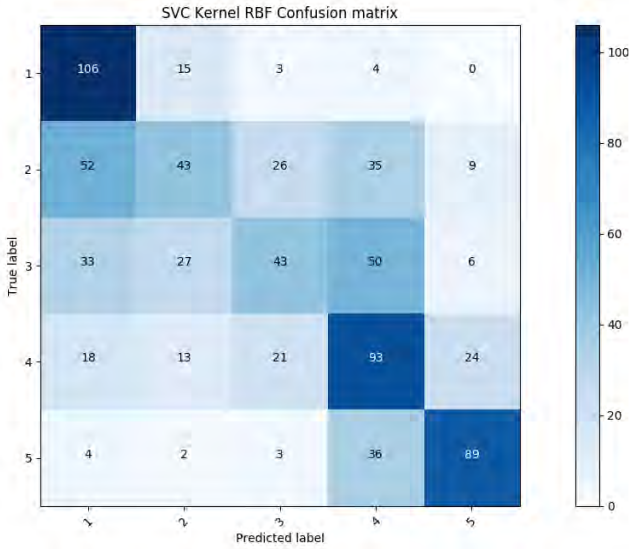


Figure 5.2: SVC Kernel RBF (Pre-Released Features)

Table 5.2: SVC Kernel RBF (Pre-Released Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	48.05	51.94	46.75	58.44	48.68	49.33	49.33	49.33	44.59	48.61	49.54

Table 5.1 shows result of each fold in 10 fold cross validation of SVC with linear kernel and their mean accuracy. Fig. 5.1 shows the confusion matrix of linear kernel where intersections of same true level and predicted level are accurate prediction of the model and all other points indicate which class the machine has predicated and what class it should be. With linear kernel, 114 out of 128 movies in class 1, 30 out of 165 for class 2, 49 out of 159

for class 3, 87 out of 169 for class 4 and 86 out of 134 movies for class 5 have been accurately classified by the machine and the mean accuracy is 48.44 percentage (Table 5.2). Next, the result of each fold of cross validation of SVC with kernel RBF are given in table 5.2. The confusion matrix (Fig. 5.2) of RBF kernel shows the predicted class like the linear one (Fig 5.1).

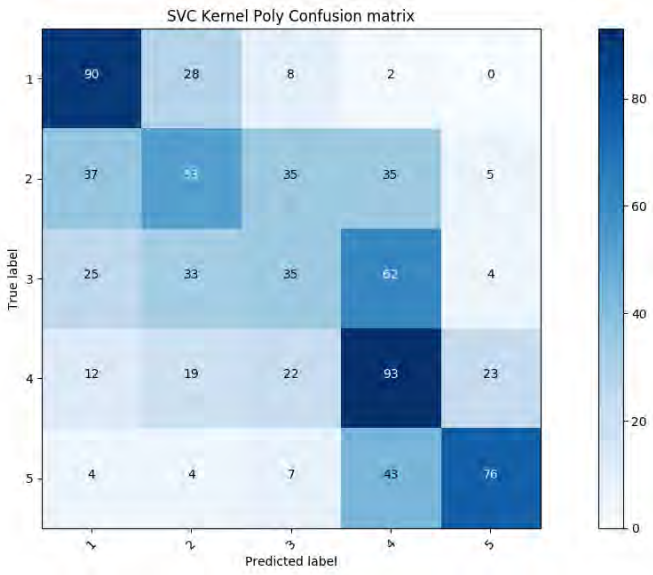


Figure 5.3: SVC Kernel Polynomial (Pre-Released Features)

Table 5.3: SVC Kernel Polynomial (Pre-Released Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	44.15	49.35	41.55	51.94	42.10	48.00	44.00	46.66	45.94	45.83	46.00

For RBF kernel, 106 out of 128 movies in class 1, 43 out of 165 for class 2, 43 out of 159 for class 3, 93 out of 169 for class 4 and 89 out of 134 movies for class 5 have been

accurately classified by the machine and the mean accuracy is 49.54 percentage (Table 5.3). Each fold result of SVC kernel polynomial are given in table 6.3. From the confusion matrix (Fig. 5.3) of SVC kernel polynomial for pre-released features, we have 90 out of 128 movies in class 1, 53 out of 165 for class 2, 35 out of 159 for class 3, 93 out of 169 for class 4 and 76 out of 134 movies for class 5 accurately classified result by the machine and the mean accuracy of 46.00 percentage (Table 5.3).

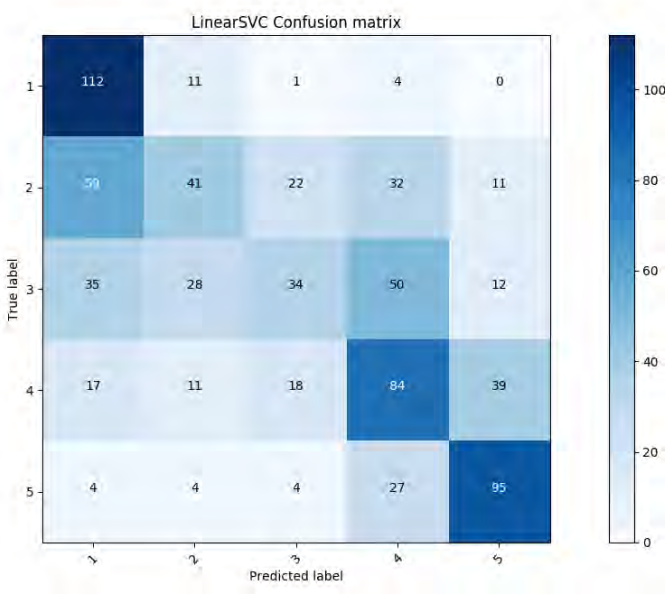


Figure 5.4: LinearSVC (Pre-Released Features)

Table 5.4: LinearSVC (Pre-Released Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	49.35	50.64	45.45	57.14	43.42	42.66	54.66	49.33	43.24	48.61	48.47

And linearSVC has 48.47 percentage of accuracy for pre-released features (Table 5.4). Its' confusion matrix (Fig. 5.4) shows the accuracy of classification where 112 out of 128 movies in class 1, 41 out of 165 for class 2, 34 out of 159 for class 3, 84 out of 169 for class 4 and 95 out of 134 movies for class 5. Table 5.4 has the accuracy in percentage for each fold of cross validation and their mean value. In this part we have merged post-released features along with pre-released features in further analysis so that after release of movies, one or few weeks later we can make a better accurate prediction.

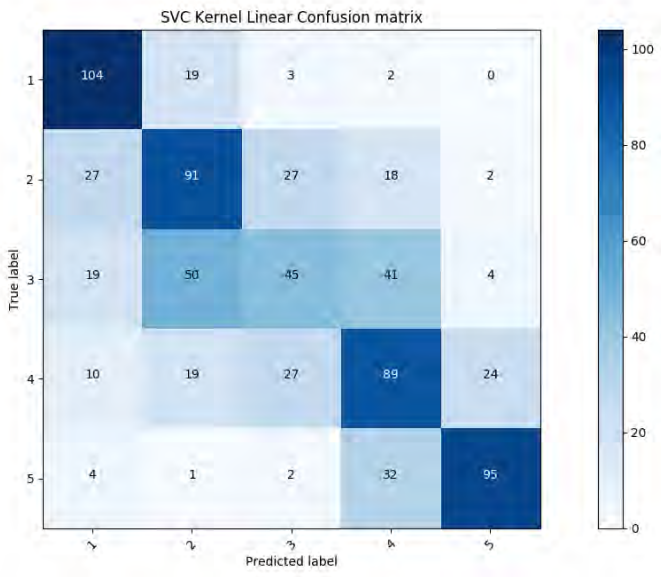


Figure 5.5: SVC Kernel Linear (All Features)

Table 5.5: SVC Kernel Linear (All Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	58.44	50.64	54.54	66.23	44.73	52.00	62.67	60.00	58.11	54.17	56.16

Then we have apply all four SVM approaches on the mixed features again. With linear kernel, 104 out of 128 movies in class 1, 91 out of 165 for class 2, 45 out of 159 for class 3, 89 out of 169 for class 4 and 95 out of 134 movies for class 5 have been accurately classified by the machine and the mean accuracy is 56.16 percentage (Table 5.5). In next, the result of each fold of cross validation of SVC with kernel RBF are given in table 5.5. The confusion matrix (Fig. 6.6) of RBF kernel shows the predicted class like the linear one (Fig. 5.5). For RBF kernel, 106 out of 128 movies in class 1, 81 out of 165 for class 2, 48 out of 159 for class 3, 102 out of 169 for class 4 and 81 out of 134 movies for class 5 have been accurately classified by the machine and the mean accuracy is 55.36 percentage (Table 5.6). Each fold result of SVC kernel polynomial are given in table 5.7. From the confusion matrix (Fig. 5.7) of SVC kernel polynomial for all the features, we have 93 out of 128 movies in class 1, 104 out of 165 for class 2, 31 out of 159 for class 3, 98 out of 169 for class 4 and 71 out of 134 movies for class 5 accurately classified result by the machine and the mean accuracy of 52.58 percentage (Table 5.7).

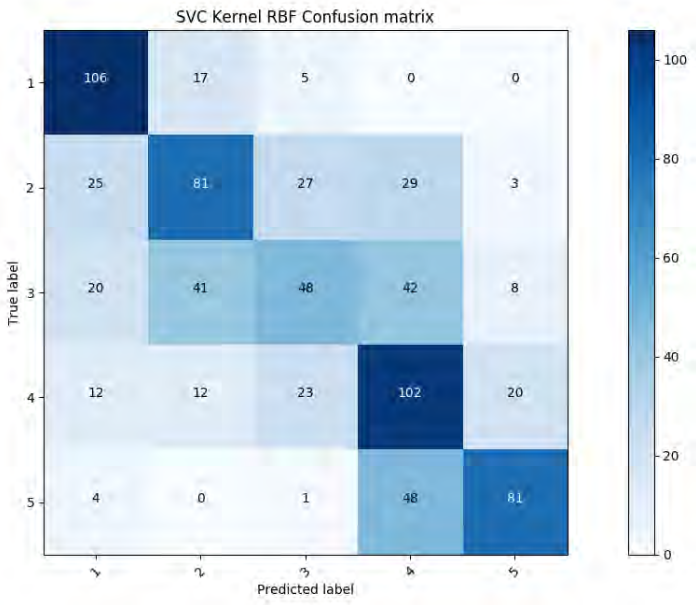


Figure 5.6: SVC Kernel RBF (All Features)

Table 5.6: SVC Kernel RBF (All Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	53.24	59.74	44.15	64.93	47.36	53.33	58.67	53.33	59.46	59.22	55.36

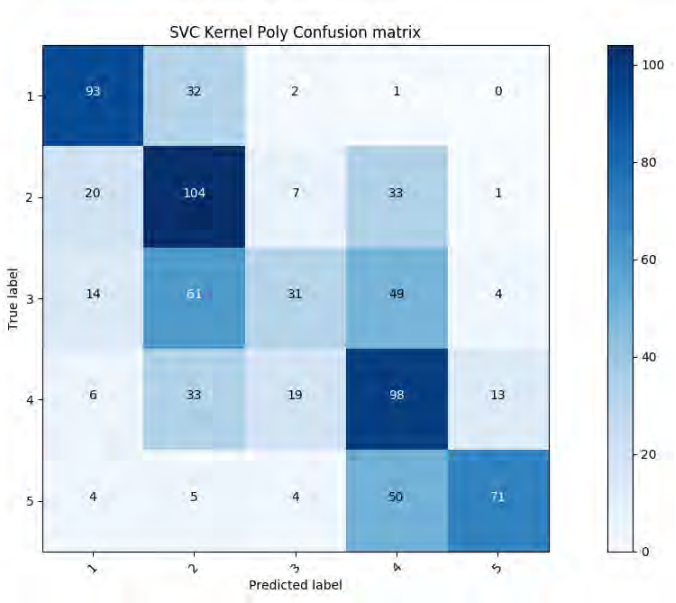


Figure 5.7: SVC Kernel Polynomial (All Features)

Table 5.8: SVC Kernel Polynomial (All Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	51.94	55.84	42.85	61.04	42.10	53.33	57.33	58.66	56.76	45.83	52.58

And linearSVC has 53.64 percentage of accuracy for all the features (table 5.8). Its' confusion matrix (Fig. 5.8) shows the accuracy of classification where 106 out of 128 movies in class 1, 83 out of 165 for class 2, 40 out of 159 for class 3, 67 out of 169 for class 4 and 109 out of 134 movies for class 5. Table 5.8 has the accuracy in percentage for each fold of cross validation and their mean value. SVM has highest exact accuracy of 49.54 percent for pre-released features and 55.36 percent for both pre-released and post-released mixed features. We also calculated one away accuracy for the model. In one away calculation, we actually consider distance of predicted class by the model from the true class. The machine has classified some movies in incorrect class for class margin values that is reason of taking consideration of one away from true class. Table 5.9 shows accuracy results of one away prediction for different SVM kernels and it is the merged output of exact classification and one class away from true one along with exact predictions. Table 4.9 shows exact vs 1 Away results accuracy in Percentage.

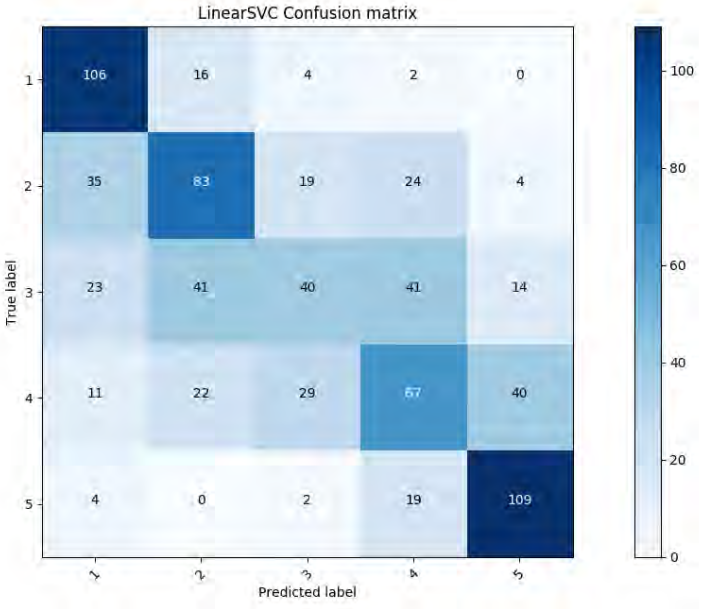


Figure 5.8: LinearSVC (All Features)

Table 5.8: LinearSVC (All Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Result(Percentage)	57.14	57.14	45.45	67.53	43.42	41.33	58.66	58.66	54.05	52.77	53.64

Table 5.9: Exact vs 1 Away Accuracy in Percentage

Kernel	Exact(Pre Released)	Exact(All)	1 Away(Pre Released)	1 Away(All)
Linear	48.44	56.16	82.11	88.87
RBF	49.54	55.36	82.25	87.54
Polynomial	46.00	52.58	83.44	85.82
LinearSVC	48.47	53.64	82.12	85.43

2D plotting is a good way for visualization. We have plotted budget vs number of screen (Fig.5.9), star power vs director (Fig. 5.10), IMDb vs Rotten Tomato reviews sentiment values (Fig. 5.11), Rotten Tomato critics rating vs audiences rating (Fig. 5.12) and Rotten Tomato critic vs audiences meter (Fig. 5.13) to understand the vector regions and data relations. From figure (Fig. 5.9, 5.10, 5.11, 5.12, 5.13), it is cleared that data are overlapping on each other which is the reason behind SVM cannot make more accurate results.

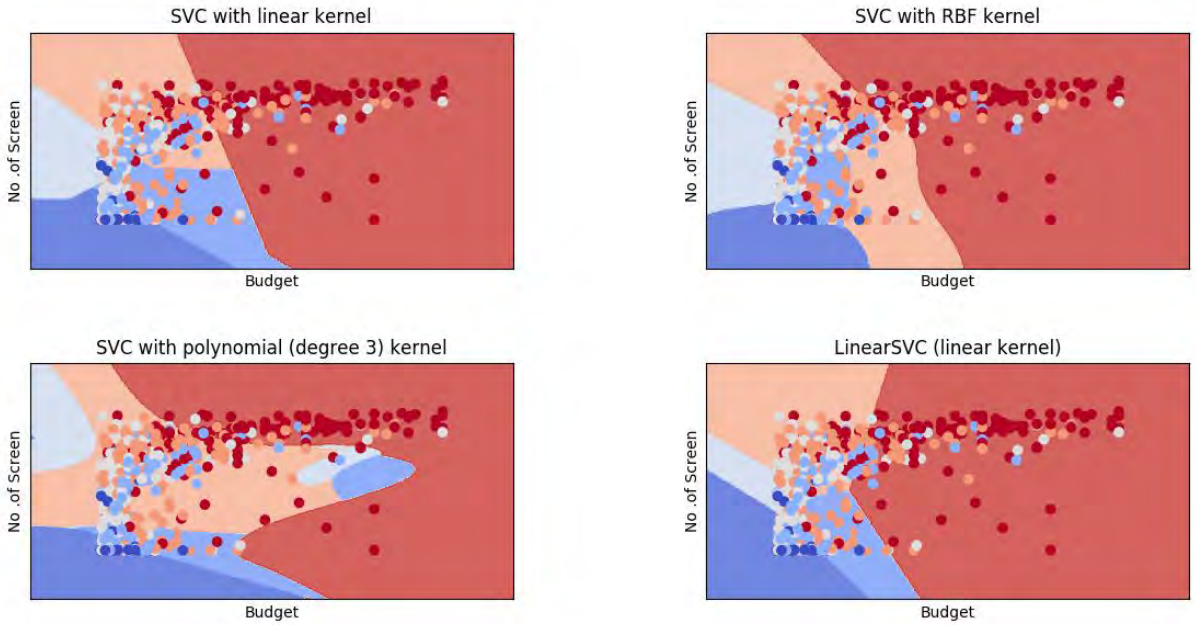


Figure 5.9: Budget vs No. of Screens

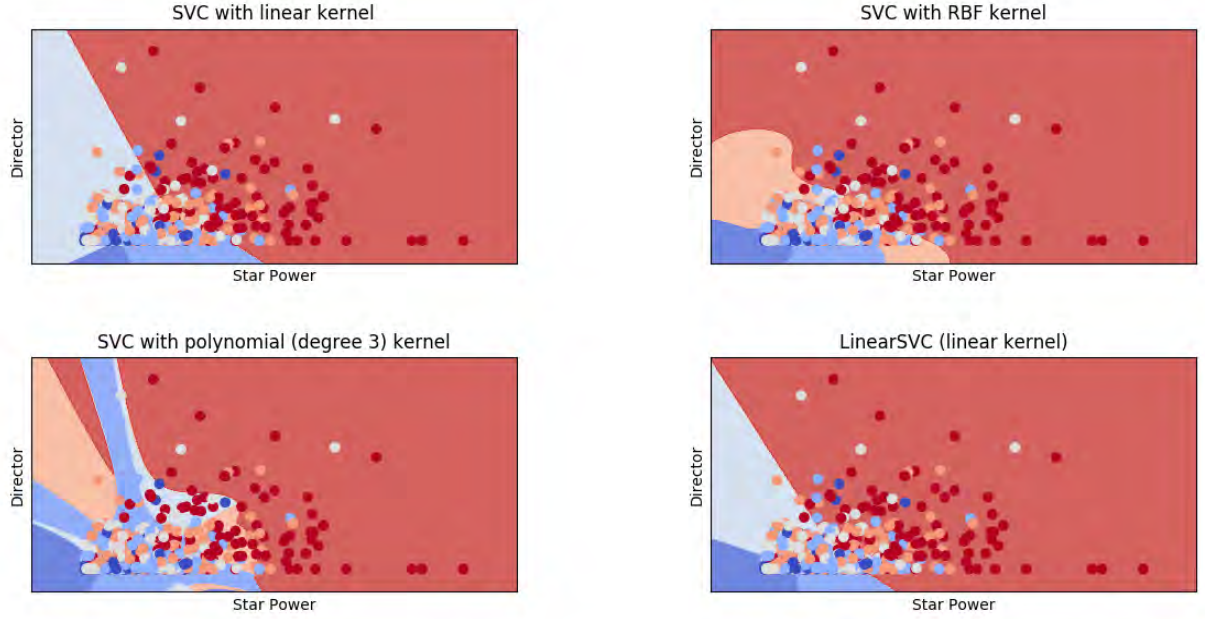


Figure 5.10: Actors/Actress vs Director Star Power

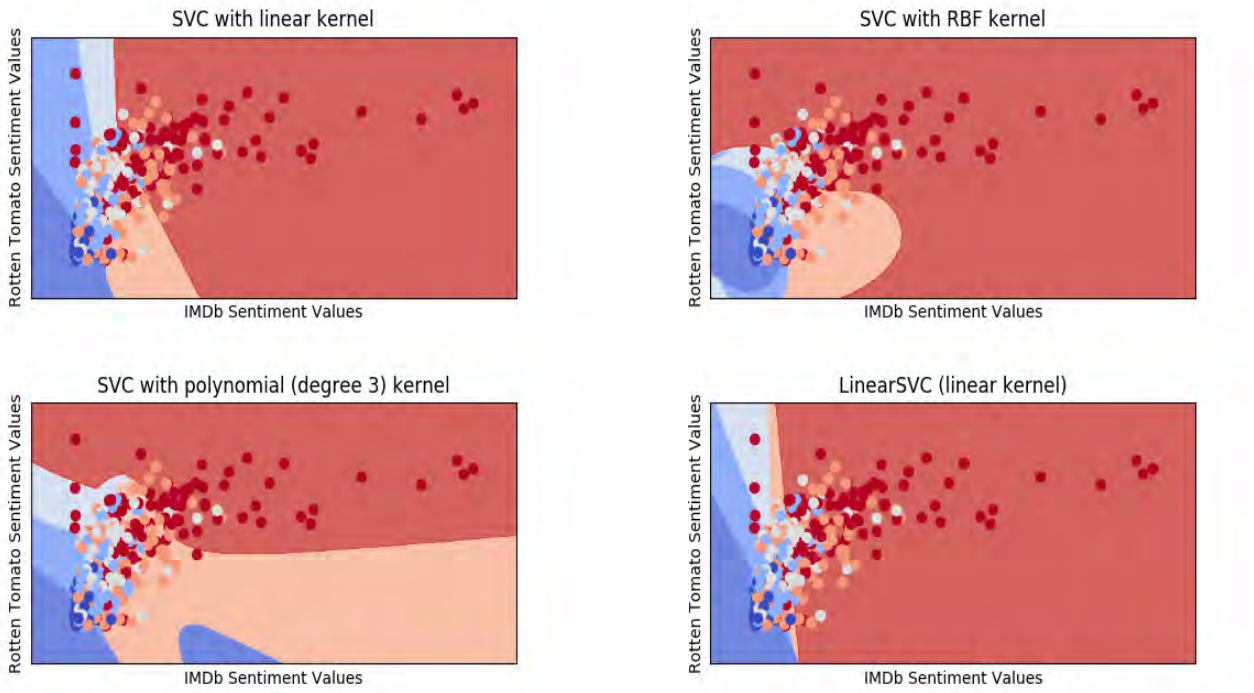


Figure 5.11: IMDb vs Rotten Tomato Sentiment values

When data are overlapping SVM is unable to make hyperplanes properly. Moreover when data is not linearly separated, data overlapping occurs and this is the main reason for SVM making comparatively poor accuracy. Fig 5.14 and Fig. 15 show the comparison exact and one away result of SVM for pre-released and all features. So our next approach is applying a neural network for better accuracy in prediction.

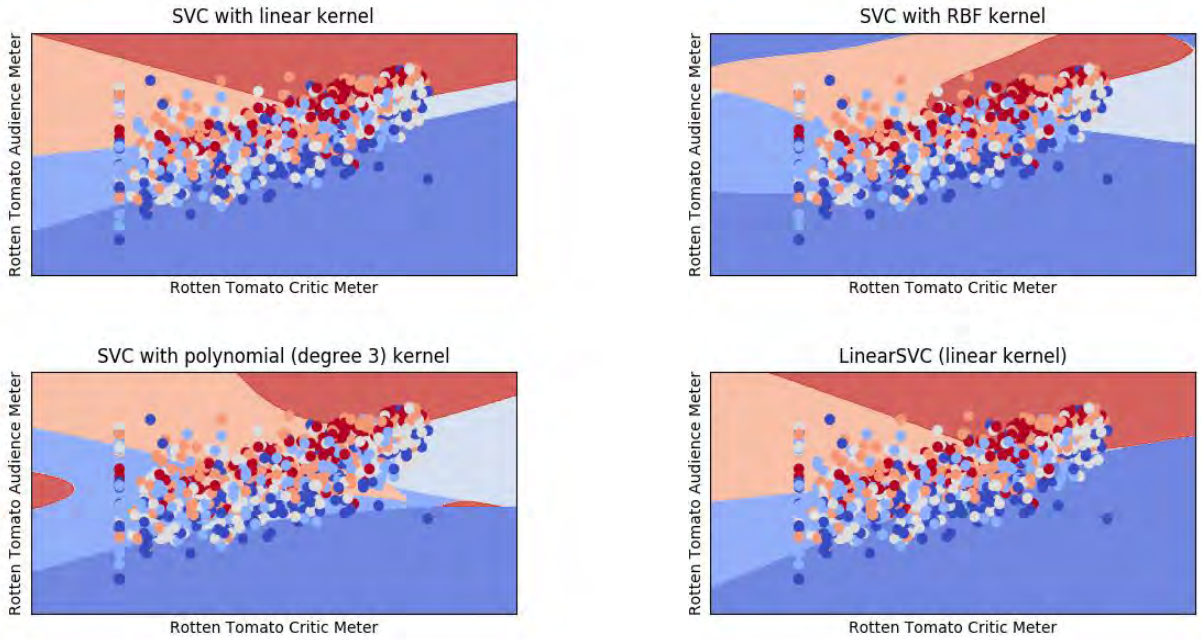


Figure 5.12: Rotten Tomato Critics vs Audience Meter

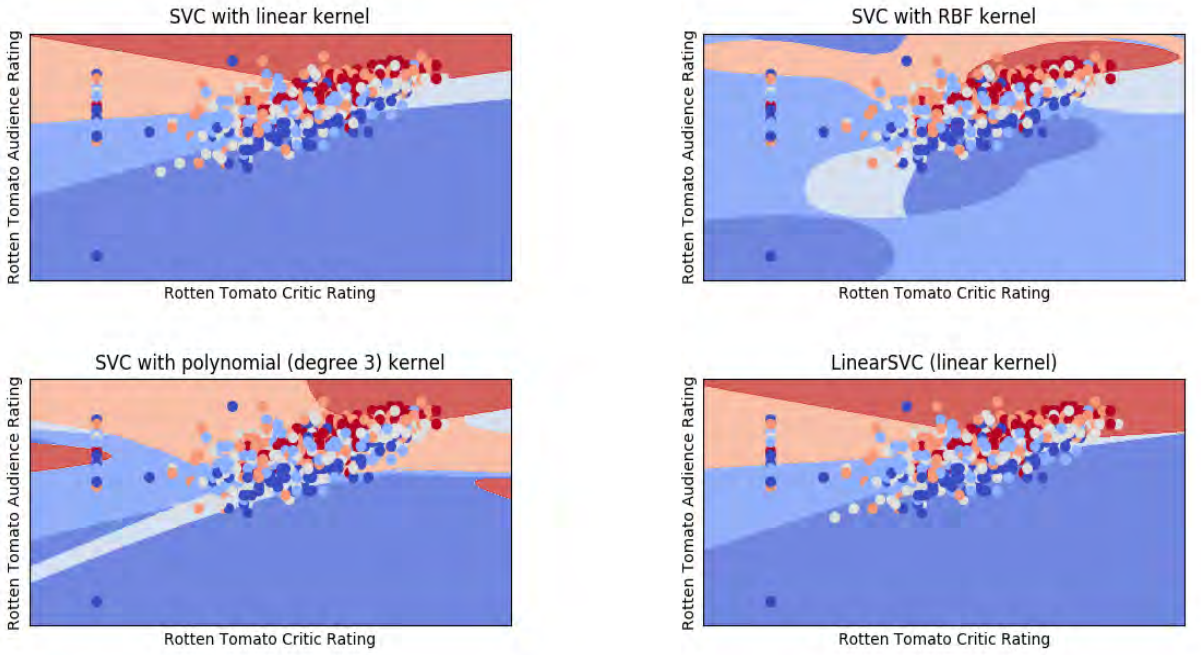


Figure 5.13: Rotten Tomato Critics vs Audience Rating

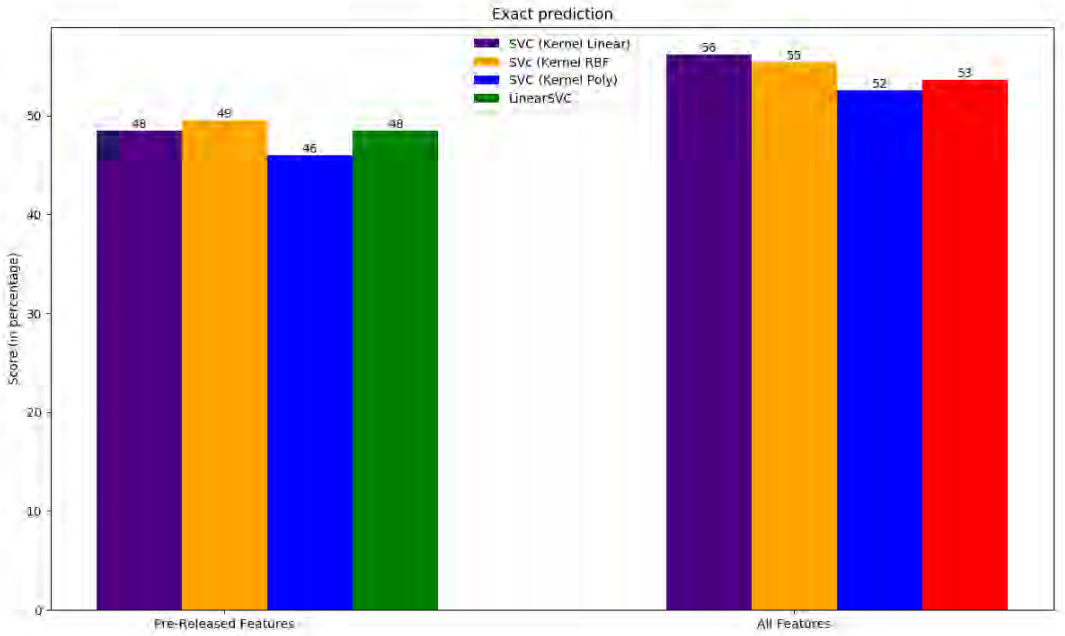


Figure 5.14: Exact Prediction

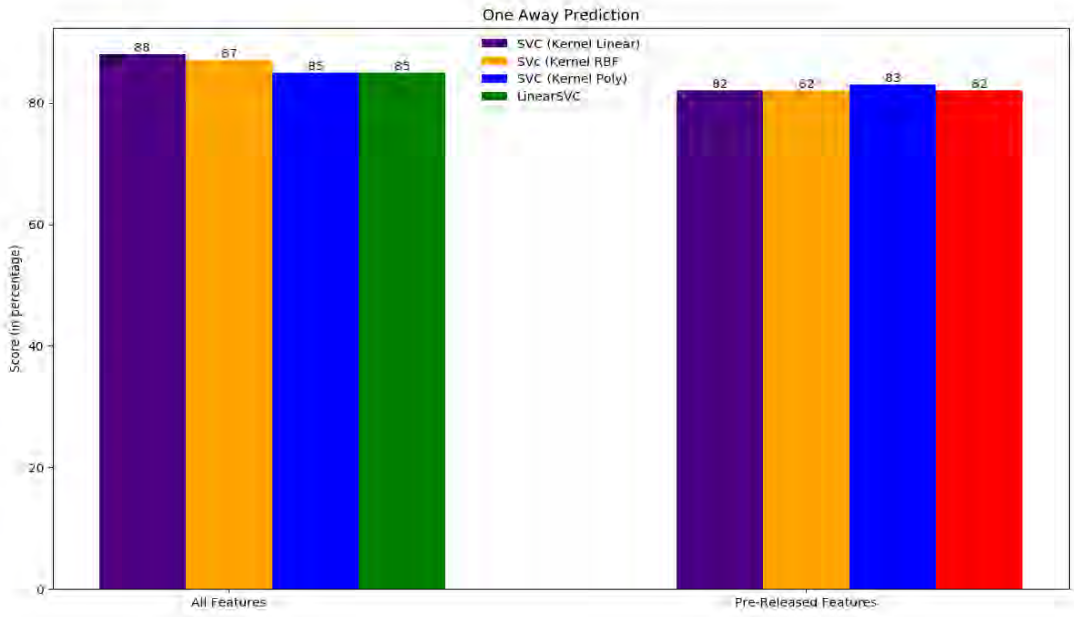


Figure 5.15: One Away Prediction

5. 3. Neural Network Analysis:

A Multi-Layer Perception Neural Network (MLP) has been used for prediction. This MLP model is developed using Keras [22], a very famous python API for neural networks. Keras sequential model has been used to build the model. Scikit-Learn [21] is also used for k-fold cross validation. In the proposed model there are three hidden layers, each has sixteen neurons. Input layer has fifteen nodes and final layer has five nodes for five outputs.

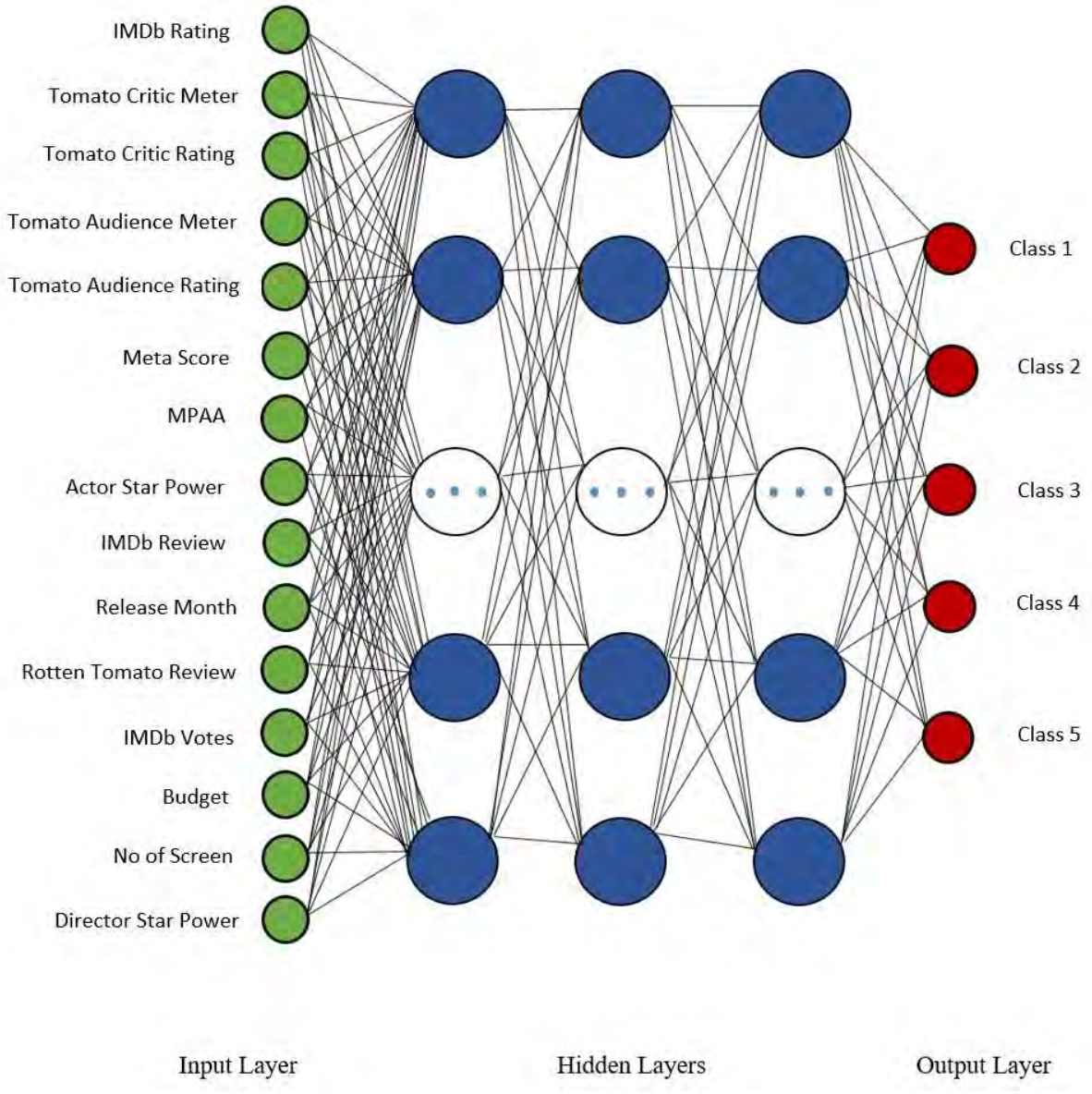
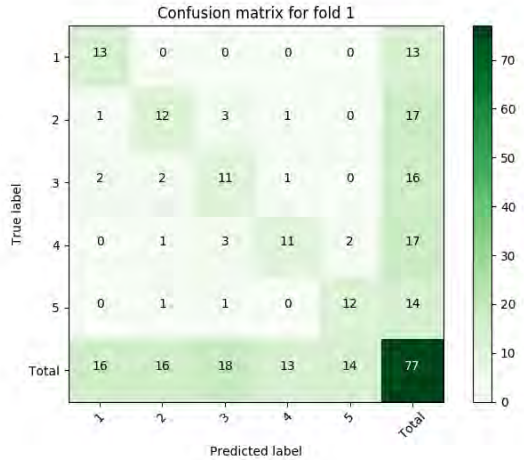
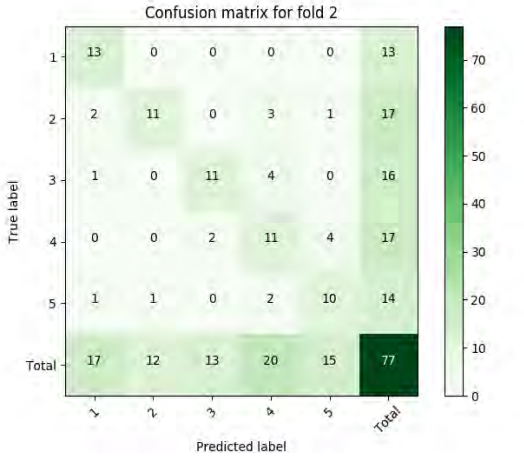


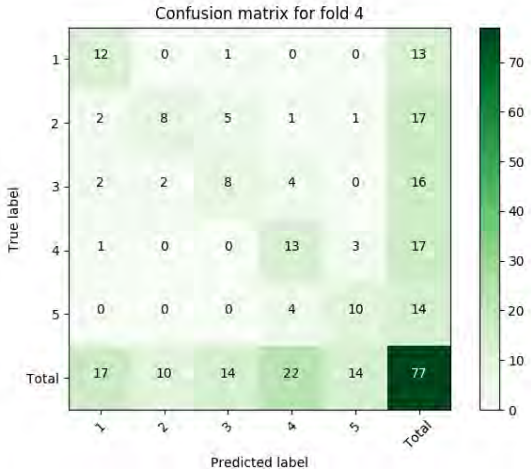
Figure 5.16: Multi-Layer Perception Neural Network

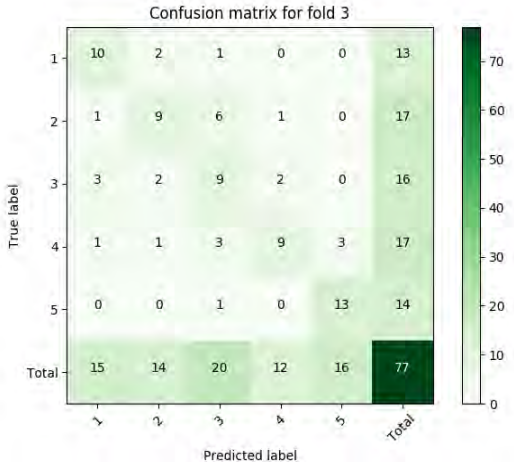
By experimenting different number of hidden layers we recognized that three hidden layer MLP architecture gives better result consistently. Fig. 5.16 represents our MLP Neural Network model. For hidden layers all nodes are not shown in this Fig 5.16 to avoid unnecessary complexity ([7]). Softmax and ReLu activation functions are used in this model for final layer and hidden layers accordingly. Overfitting is a common problem in Neural Networks. Overfitting causes very small training error but when a new test set comes for prediction, error increases very much because the network fits too much on training examples, as a result it gives poor prediction accuracy for new test data. For overfitting problem dropout regulation has been inserted after each hidden layer. Dropout is a very common and good solution for overfitting, it drops out random neurons during training to avoid overfitting.

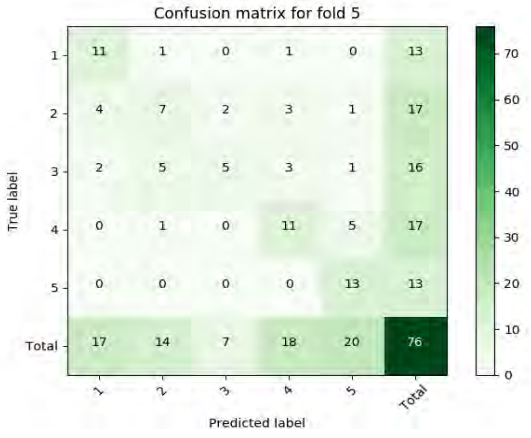
Neural Networks learns through changing weights in a direction to minimize loss. Learning rate is important to fit a model perfectly. Too small learning rate takes months to train the whole network where too large learning rate causes under fitting the network and a huge loss during training. There are several optimizers available in keras like SGD, RMSprop, Adagard, Adam etc. Among these optimizers Adam and Adagard can handle learning rate on their own. In this model, Adam has been used as an optimizer. So when some arbitrary neurons are dropped out, other neurons will try to make the prediction for missing neurons, and this way the network will be better generalized. The total network was trained and tested using k-fold cross validation technique. Which is best for testing because it ensures that there are no bias selection of test and train data.To predict upcoming movies we included some pre-release feautures. When our MLP model has been trained using only pre-release features we got 68% accuracy and 90.2% accuracy if we consider one away prediction. For confusion matrices of each fold for only pre-release features are given in Fig. 5.17. Fig. 5.18 shows final confusion matrix after 10 fold cross validation.

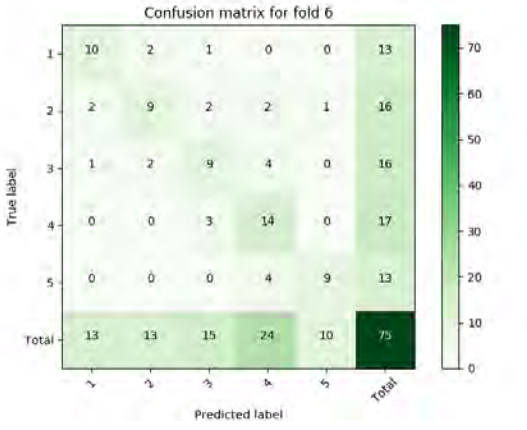

Confusion matrix for fold 1
True label
Predicted label
1 2 3 4 5 Total
13 0 0 0 0 13
1 12 3 1 0 17
2 2 11 1 0 16
0 1 3 11 2 17
0 1 1 0 12 14
16 16 18 13 14 77
0 10 20 30 40 50 60 70


Confusion matrix for fold 2
True label
Predicted label
1 2 3 4 5 Total
13 0 0 0 0 13
2 11 0 3 1 17
1 0 11 4 0 16
0 0 2 11 4 17
1 1 0 2 10 14
17 12 13 20 15 77
0 10 20 30 40 50 60 70


Confusion matrix for fold 4
True label
Predicted label
1 2 3 4 5 Total
12 0 1 0 0 13
2 8 5 1 1 17
2 2 8 4 0 16
1 0 0 13 3 17
0 0 0 4 10 14
17 10 14 22 14 77
0 10 20 30 40 50 60 70


Confusion matrix for fold 3
True label
Predicted label
1 2 3 4 5 Total
10 2 1 0 0 13
1 9 6 1 0 17
3 2 9 2 0 16
1 1 3 9 3 17
0 0 1 0 13 14
15 14 20 12 16 77
0 10 20 30 40 50 60 70


Confusion matrix for fold 5
True label
Predicted label
1 2 3 4 5 Total
11 1 0 1 0 13
4 7 2 3 1 17
2 5 5 3 1 16
0 1 0 11 5 17
0 0 0 0 13 13
17 14 7 18 20 76
0 10 20 30 40 50 60 70


Confusion matrix for fold 6
True label
Predicted label
1 2 3 4 5 Total
10 2 1 0 0 13
2 9 2 2 1 16
1 2 9 4 0 16
0 0 3 14 0 17
0 0 0 4 9 13
13 13 15 24 10 75
0 10 20 30 40 50 60 70

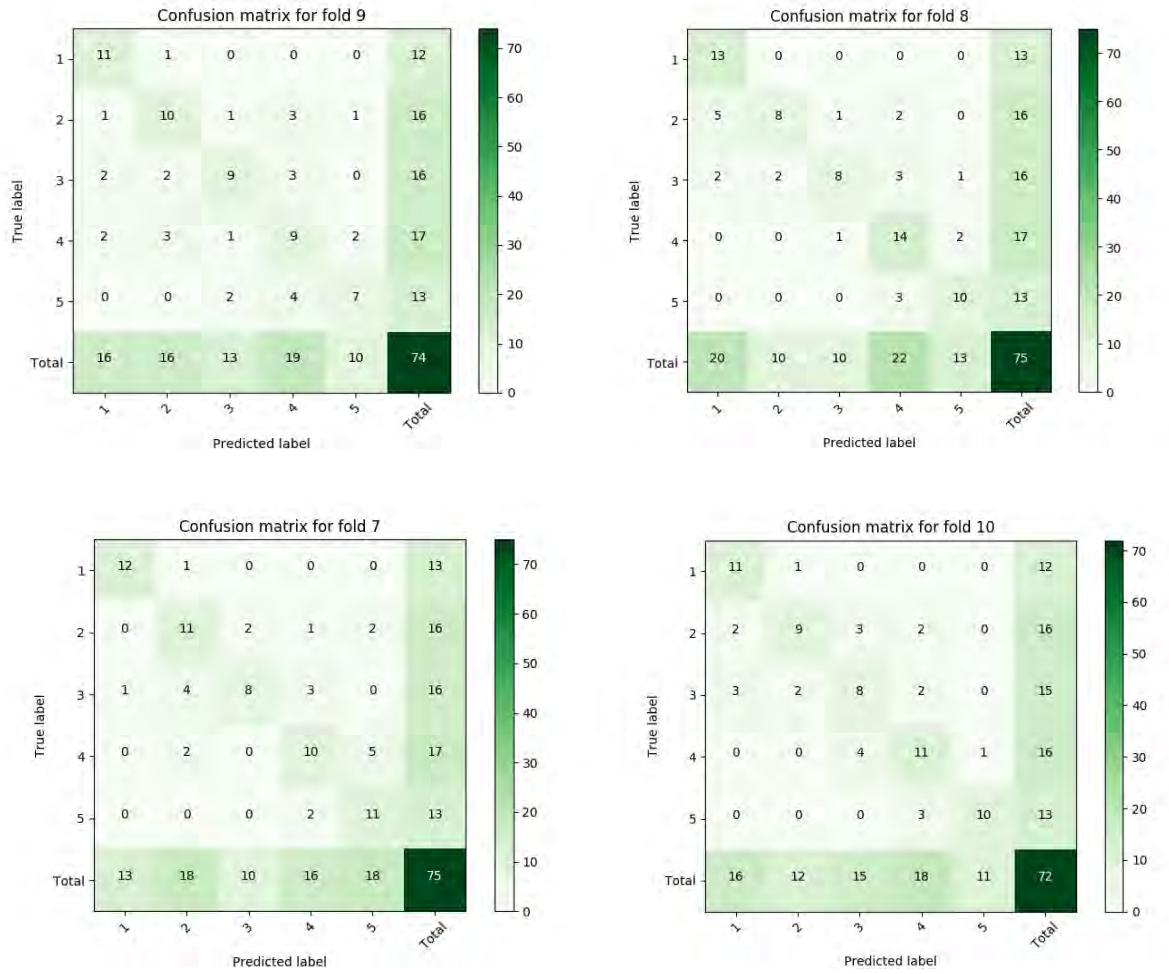


Figure 5.17: Each Fold of 10 Fold Cross Validation (Pre-Released Features)

Table 5.10 shows the performance of Neural Network while trained with only pre-release features .For each fold exact prediction and considering one away prediction accuracy has shown. With only pre-release features included Neural Network exactly predicts 68% of all movies. When one away prediction is considered, accuracy goes to 88.5%.

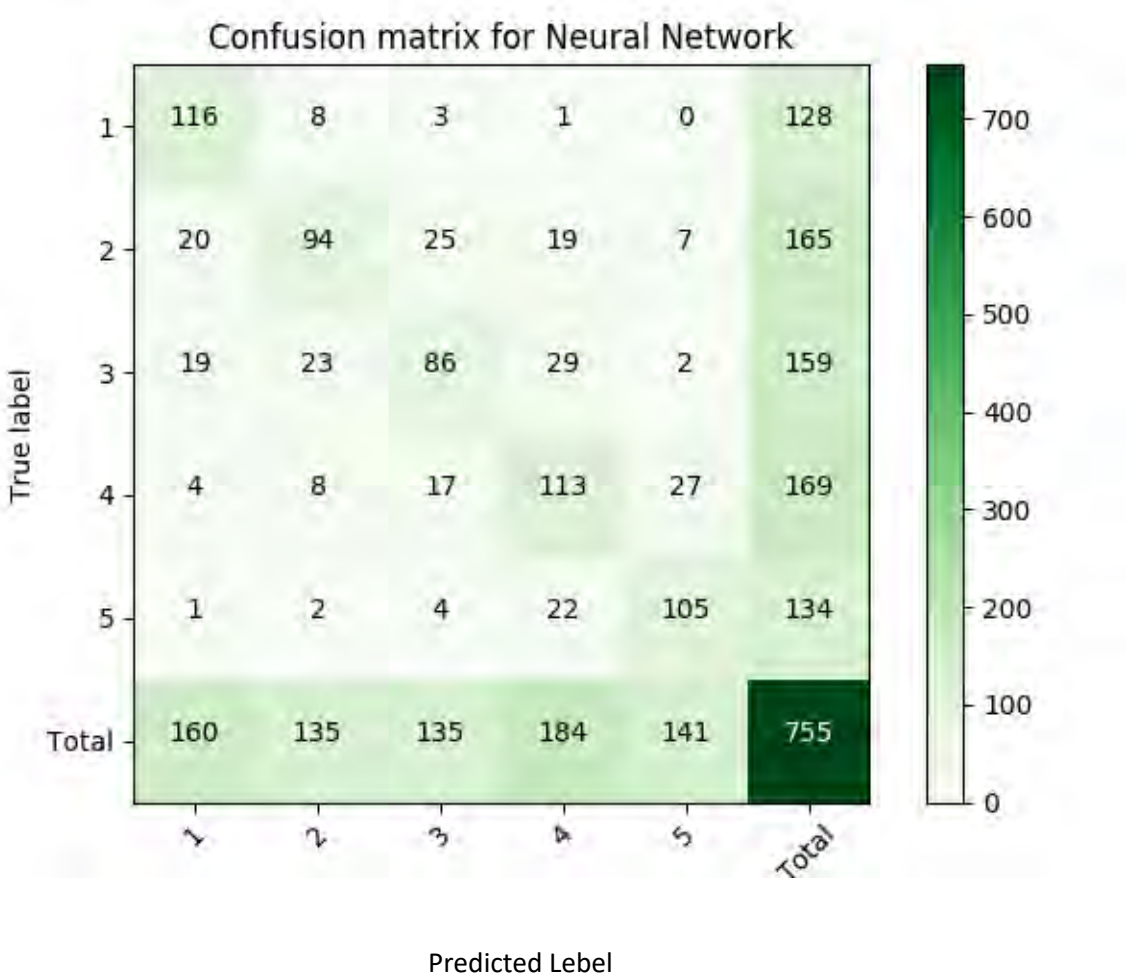
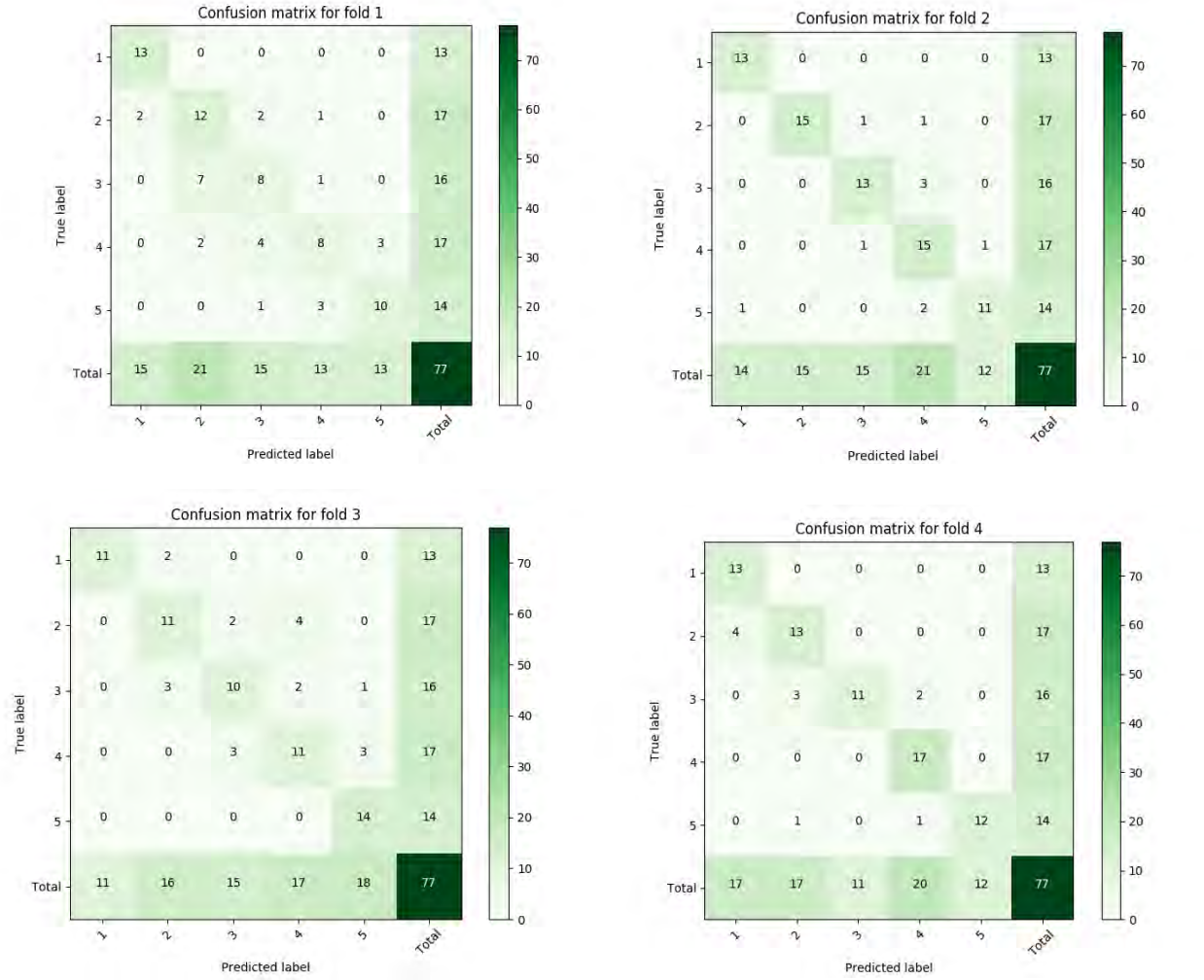


Figure 5.18: Confusion Matrix (Pre-Released Features)

Table 5.10: Performance analysis (Pre-Released Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Exact Match	76.6%	72.5%	64.9%	66.2%	61.8%	68%	69.3%	70.6%	62.1%	68%	68%
One Away	92.12%	90.6%	89.5%	92.1%	88.1%	85.3%	91.3%	82.6%	82.1%	93%	88.5%

While training with post and pre-release features, 80.2% accuracy has been achieved with 10 fold cross validation. In Fig. 5.19 confusion matrices are shown for each fold and in Fig. 5.20 a single confusion matrix has shown which includes all 10 folds.



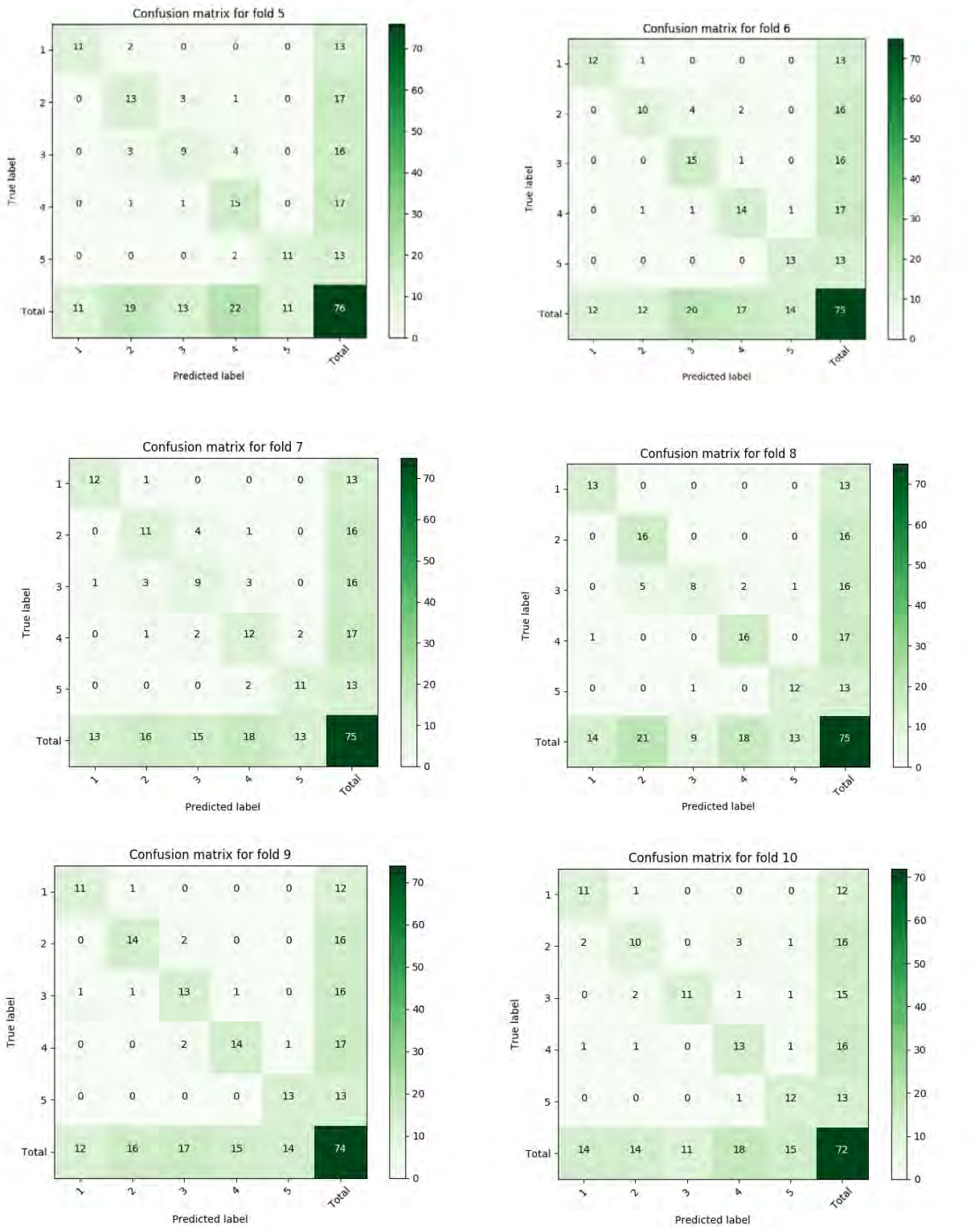


Figure 5.19: Each Fold of 10 Fold Cross Validation (All Features)

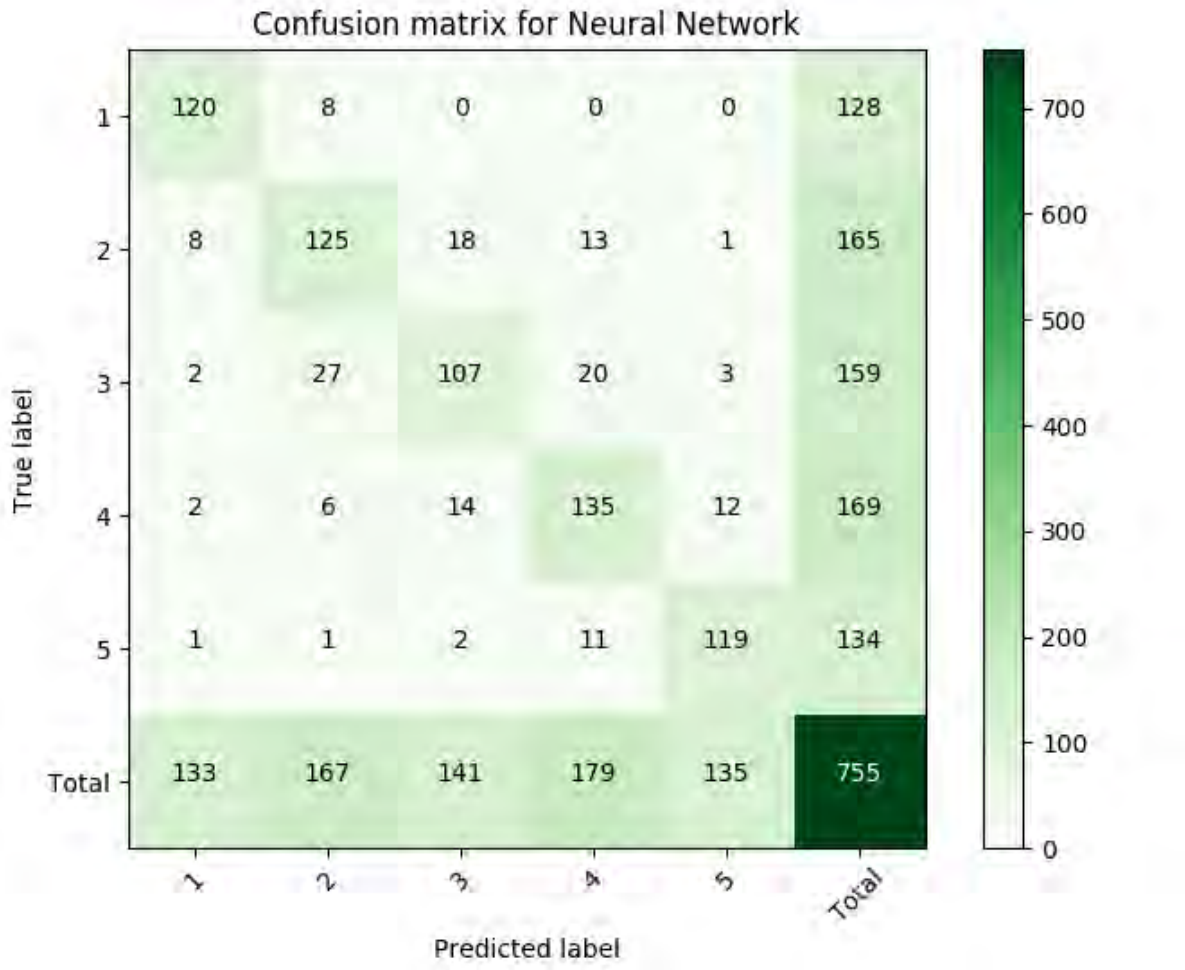


Figure 5.20: Confusion Matrix (All Features)

Table 5.11: Performance Analysis (All Features)

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Exact Match	66.2%	87%	74%	85.7%	77.6%	85.3%	73.3%	86.6%	87.8%	79.1%	80.2%
One Away	94.7%	97.4%	93.4%	98.6%	97.3%	95.9%	95.9%	94.9%	98.6%	90.2%	95.6%

Table 5.11 shows the performance of Neural Network while trained with both pre and post-release features. For each fold exact prediction accuracy and accuracy considering one away prediction has shown. With all features included Neural Network exactly predicts 80.2% of all movies. When one away prediction is considered, accuracy goes to 95.6%.

From our analysis, we have figured out which attributes have more effect in making the prediction. Importance of pre-released features are shown in Fig. 5.21 and all features are shown in Fig. 5.22.

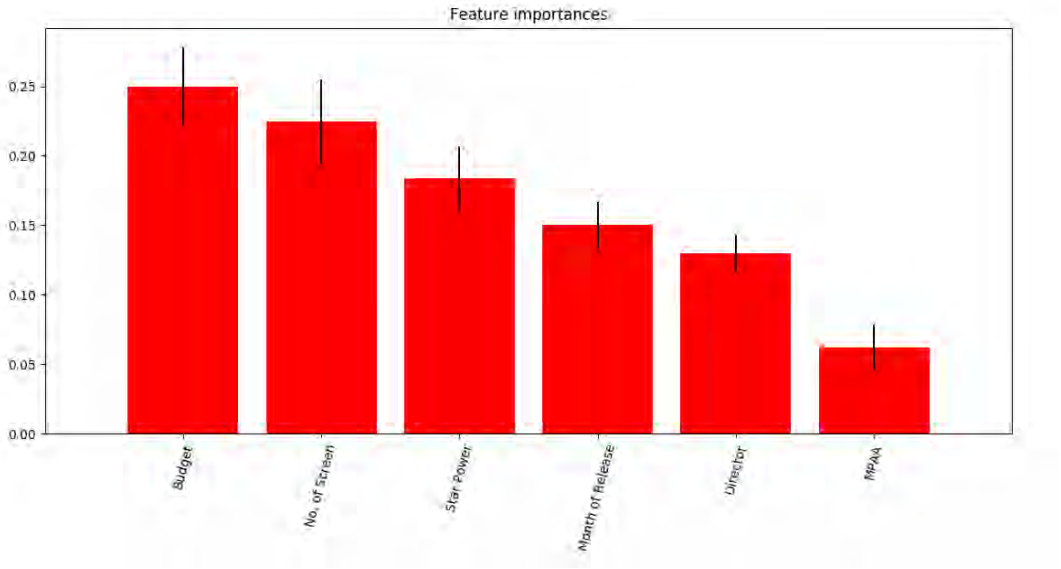


Figure 5.21: Importance of pre-released features

As a pre-released feature, budget is the most important one among all pre-released and all other post-released features, which is not surprising (Fig. 5.21 and 5.22). More budget means more star actor/actress, more marketing and so on. And then we can see in Fig. 5.21 that number of screens is the second most important pre-released feature. For all features it is the third important one (Fig. 5.22). In Fig 5.23 we can see that number of screen consistently increases when class increases. Higher class value means more successful movie in terms of profit. Fig. 5.23 also shows that the relation between number of screen and movie profit class

is very strong.

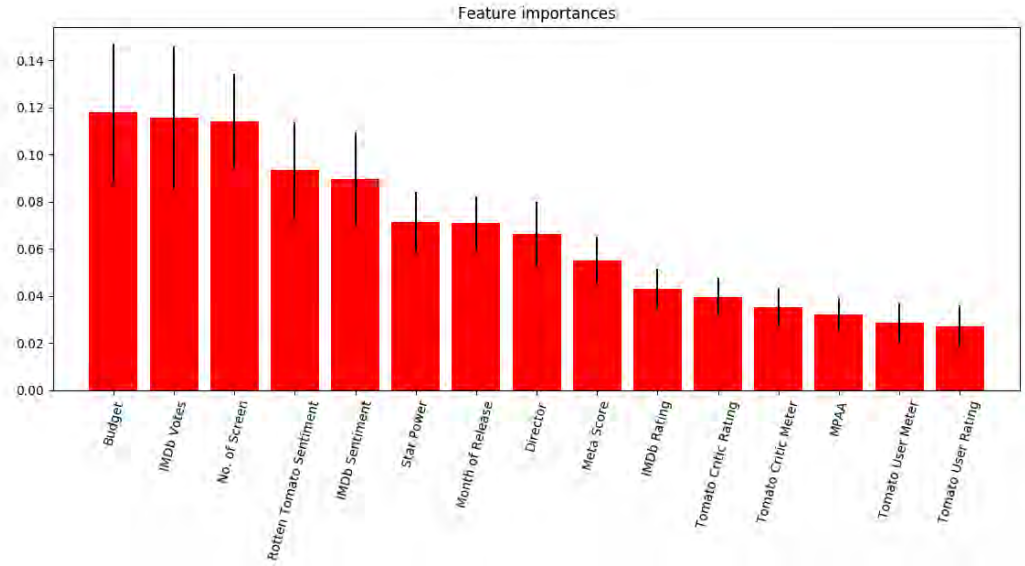


Figure 5.22. Importance of all features

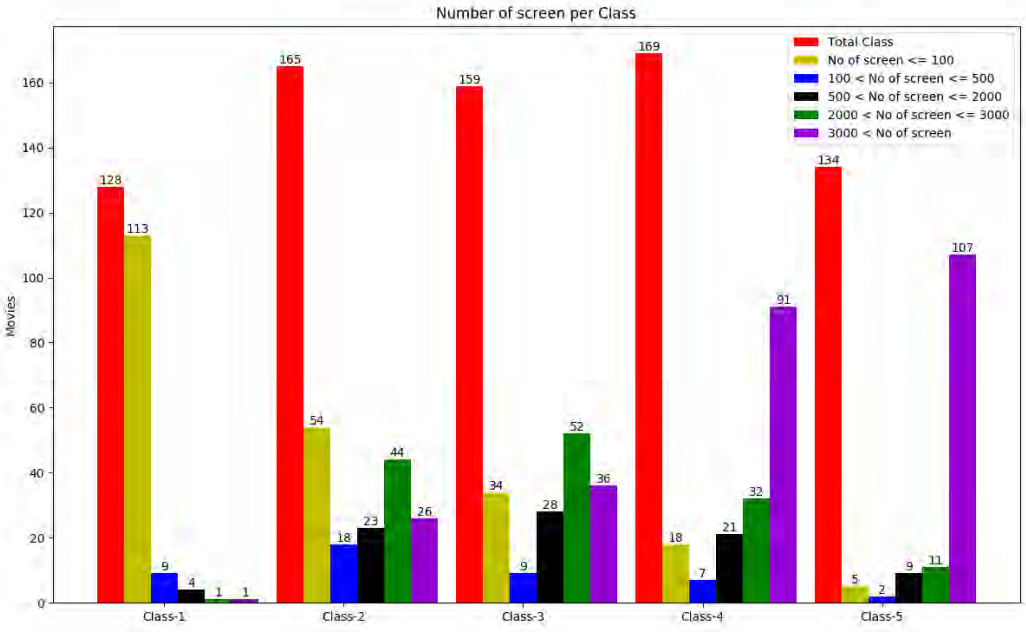


Figure 5.23: Relation between Target Class and No. of Screens.

We also figured out that the month released is another important feature. When a movie is going to release is an important fact, for example if a movie releases during Christmas, the probability of good amount of profit for that movie goes higher. Fig. 5.24 shows a relation between release month and movie target classes. Among all pre-released features, directors' star power and MPAA have lower effect than other features in the predictions (Fig. 5.21)

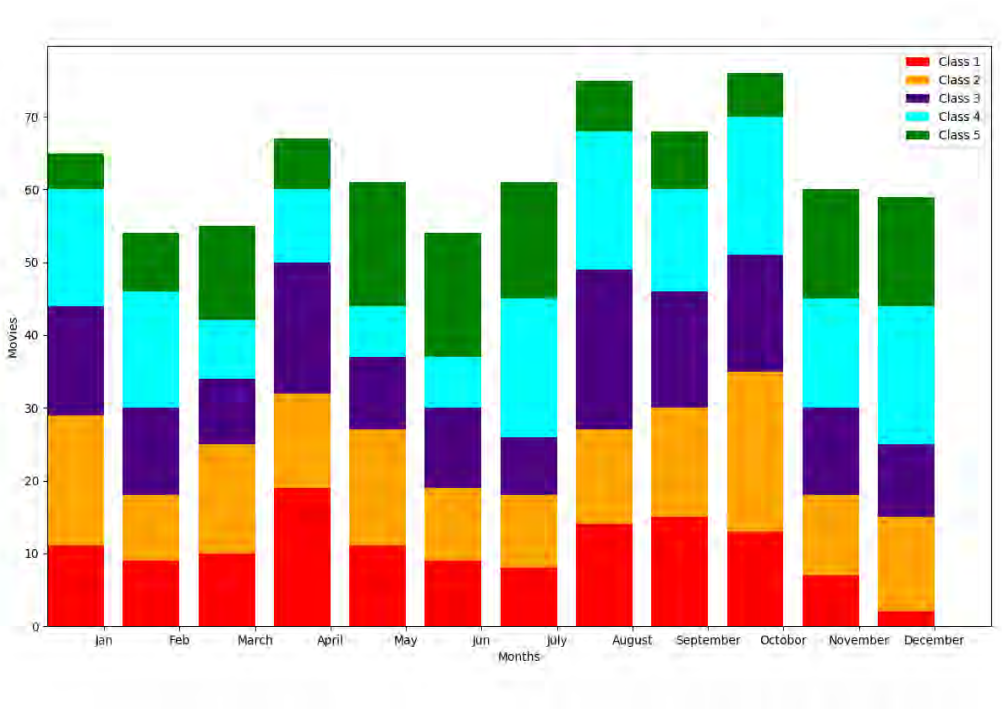


Figure 5.24: Relation between Target Class and Month of Release

In Fig. 5.25 a performance comparison between SVM and Neural Network has been shown. For each class, the number of movies has been exactly predicted for both pre-released and all features, are visualized in this comparison.

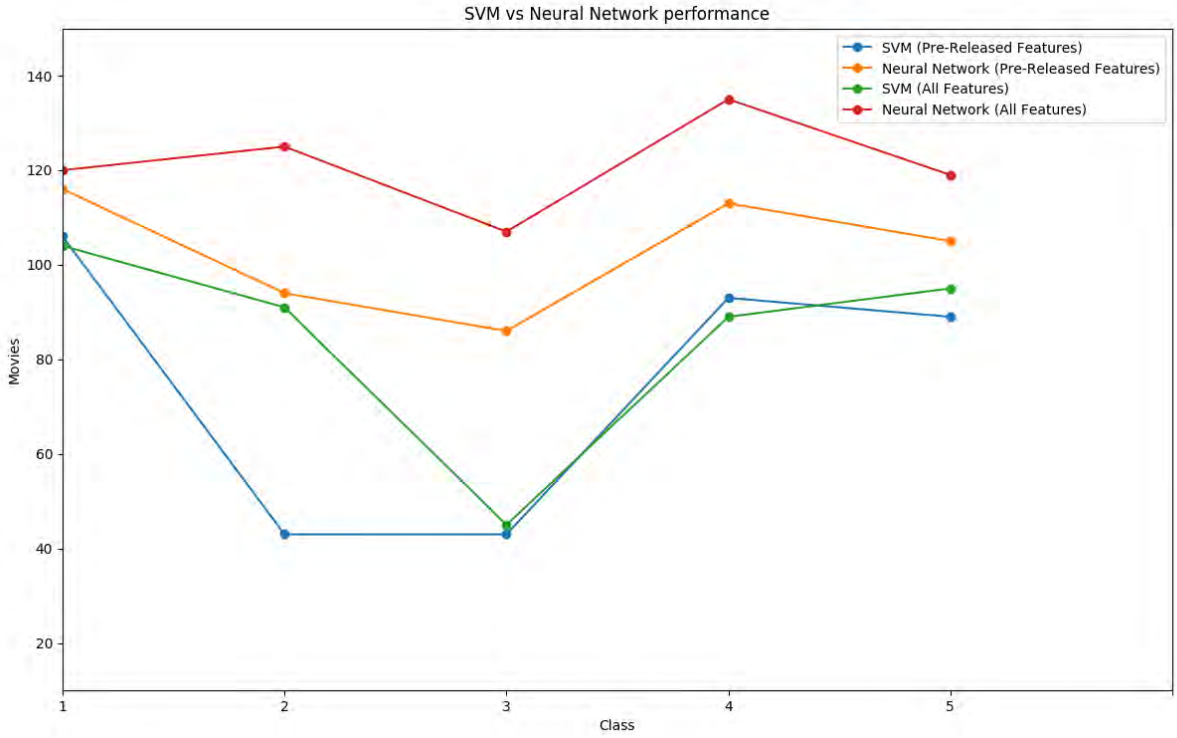


Figure 5.25: Performance Comparison between SVM and Neural Network

6. Conclusion and Discussion

6. 1. Discussion and Future Planning:

A movie success does not only depend on features related to movies. Number of audience plays very important role for a movie to become successful. Because the whole point is about audiences, the whole industry will make no sense if there is no audience to watch a movie. Number of ticket sold during a specific year can indicate the number of audiences of that year. And movie audience depend on many features like political conditions and economic stability of a country. GDP of a country can be used as a feature to know if there was economic stability during the time period when a movie released. During an economical depression, very few amount of audience will go to the theatres to enjoy movies. So these facts plays a vital role for an ultimate success of a movie. So, for future work we suggest to take these features in consideration.

6. 2. Conclusion:

We did not consider genre and sequel of a movie as a feature. Prediction of a sequel movie is tough, some movies gain a good amount of profit only for its previous sequel. Some other researches also avoided sequel [10]. Some research paper considered only pre-release features for prediction ([10], [12]), some considered mostly post-release data ([1], [2]). But in our research we considered both features to predict both upcoming and recently released movies. Support Vector Machine (SVM) exact prediction 49.54% and one away prediction accuracy 83.4% is a good score. SVM also produced good result using all features, exact prediction 56.1% and one away prediction accuracy is 88.8%. But SVM gives relatively bad result comparing to Neural Networks. For Neural Networks, 80.2% exact prediction, 95.6% accuracy with one away prediction using all features and 68% exact prediction, 88.5% one away prediction using only pre-release data is a very good score.

7. References

[1] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. D. Gregorio, "Prediction of movies box office performance using social media," *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 2013.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] R. Sharda and E. Meany, "Forecasting gate receipts using neural network and rough sets," in *Proceedings of the International DSI Conference*, 2000, pp. 1–5.

[8] 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.

[9] 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.

[10] M. T. Lash and K. Zhao, “Early Predictions of Movie Success: The Who, What, and When of Profitability,” *Journal of Management Information Systems*, vol. 33, no. 3, pp. 874–903, Feb. 2016.

[11] 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.

[12] 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.

[13] 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.

[14] M.H Latif, H. Afzal “Prediction of Movies popularity Using Machine Learning Techniques”, National University of Sceinces and technology, H-12, ISB, Pakistan.

[15] K. Jonas, N. Stefan, S. Daniel, F. Kai “Predicting Movie Success and Academi Awards through Sentiment and Social Network Analysis” University of Cologne, Pohligstrasse 1, Cologne, Germany.

[16] J. Duan, X. Ding, and T. Liu, “A Gaussian Copula Regression Model for Movie Box-office Revenue Prediction with Social Media,” *Communications in Computer and Information Science Social Media Processing*, pp. 28–37, 2015.

[17] L. Doshi, J. Krauss, S. Nann, and P. Gloor, “Predicting Movie Prices Through Dynamic Social Network Analysis,” *Procedia - Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6423–6433, 2010.

[18] T. G. Rhee and F. Zulkernine, “Predicting Movie Box Office Profitability: A Neural Network Approach,” *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016.

[19] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, “Predicting movie Box-office revenues by exploiting large-scale social media content,” *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1509–1528, Feb. 2014.

[20] Z. Zhang, B. Li, Z. Deng, J. Chai, Y. Wang, and M. An, “Research on Movie Box Office Forecasting Based on Internet Data,” *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, 2015.

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

[22] F. Chollet, Keras (2015), GitHub repository, https://github.com/fchollet/keras

[23] B. R. Litman & H. Ahn (1998). Predicting financial success of motion pictures. In B. R. Litman (Ed.), the motion picture mega-industry. Boston, MA: Allyn & Bacon Publishing, Inc.

[24] J. Valenti (1978). Motion Pictures and Their Impact on Society in the Year 2000, speech given at the Midwest Research Institute, Kansas City, April 25, p. 7.