iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** IV **Month of publication:** April 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Success Prediction of Films at Box Office Using Machine Learning

Parag Ahivale[1], Omkar Acharya[2]
[1,2]*Pune Institute of Computer Technology*

Abstract— Feature films are a multi-billion industry. Here prediction of a movie's success is predicted based on its features like cast, genre of movie, month of release, run time, directors, producers etc. Based on multiple such features and with the database of previous movies statistics, machine learning algorithm like Linear Regression can predict the approximate ratings the movie can receive once it is actually released and hence classify a movie as a hit or a flop. A large amount of data representing feature films is maintained by the Internet Movie Database (IMDb). Many of the movies listed on IMDb contain an average user rating on a scale of 0 to 10 which corresponds to public opinion of that movie. Due to the large number of films produced as well as the level of scrutiny to which they are exposed, it may be possible to predict the success of an unreleased film based on publicly available data This data can be extracted and prepared for use in training machine learning algorithms .The goal of this paper is to discuss a system that can closely predict average user rating by learning from historical movie data and hence determine if it is likely to be a flop or a hit.

Keywords— Linear Regression Technique, Logistic Regression Technique, Movie Database, Movie Ratings, Prediction

I. INTRODUCTION

In India 1000s of films are released every year. Cinema in India is a multi-million industry where even some individual films earn over 50 million rupees. Large production houses control most of the film industry, with billions of rupees spent on advertisements alone. Given the sheer number of films produced as well as the level of scrutiny to which they are exposed, it may be possible to predict the success of an unreleased film based on publicly available data. A large amount of data representing feature films is maintained by the Internet Movie Database (IMDb). If it was somehow possible to know beforehand the likelihood of success of the movies, the production houses could adjust the release of their movies so as to gain maximum profit. They could use the predictions to know when the market is dull and when it's not. In recent years, various sentiment analysis techniques were successfully applied to analysing user reviews, which in turn were applied to predict movie ratings. This shows a dire need for such software to be developed. Many have tried to accomplish this goal of predicting movie success. None of the studies thus far has succeeded in suggesting a model good enough to be used in the industry. In this study, an attempt to use historic movie data to predict the rating for a film is made. Here Linear Regression with multiple variables is used to predict the rating of the films and thus if it will be a hit/flop. Movie data of about 500 past movies was used to train the learning algorithm. Selecting features of movie to train the algorithm is a challenging part as one may feature may have a big impact in some cases while none in the other. Instead of selecting all the features of a movie, only seven most crucial features were selected as a part of data pruning. These seven features: producer, director, lead actor, lead actress, genre, runtime, month of release as they tend to make a big impact on the success of a film at the Box Office. Sophisticated implementation of such a model will help audience to make better choices as well help production houses to come up with better combination of cast, genre etc. Also such a model will also help them to plan a successful date of release and an optimum runtime so as to keep the audiences eye-balls glued to the screen.

II. RELATED WORK

A. Linear Regression

The basic form of a k-variable linear regression model is defined as follows:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$

Variable y is known as the prediction, variables x_i, i = 1, ..., k are the features and ε represents the error term.

The linear regression model written in matrix form:

$Y = X\beta + \varepsilon$

B. LOGISTIC REGRESSION

The logistic regression function is defined as:

$h_\Theta(x) = g(\theta^T x)$

Where θ is the weight vector learned and g( ) is a sigmoid function.

III. PROPOSED MATHEMATICAL MODEL

Our goal is to predict the ratings of new movie based only on the ratings of the training set examples. Both linear regression and logistic regression are well-known methods. Their use is motivated by the limited number of examples we have for training the predictors. Being simple models, they can be expected to work better in these conditions than more complex models, such as Support Vector Regression, which generally require more training samples. Due to its simplicity and widespread use in machine learning applications, linear regression can be considered the reference method here.

We have considered 500 training examples, each of which has seven features and the corresponding human ratings of some particular attribute. We then form a prediction model that can be used to predict the human rating when the computational features are known.

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n + \varepsilon$$

The goal is to define a relationship between the prediction value and the features by solving for the linear coefficients, θ that best map the features to the prediction value. Where, the ratings have been collected in a vector Y. Y is a (m x 1) vector. In our case m=500.

INPUT: Training set X and Y. In this case training set of 500 examples has been aggregated. Along with this the input x for which rating is to be calculated is also a part of input.

OUTPUT: Rating of the movie h(x)

Regularized Cost Function: $Jcost(\Theta) = 1/2*m*sum(X^T\Theta - y)^2 + \lambda/2/m*sum((\Theta)^2)$

Gradient descent to set parameter Θ: $\Theta = \Theta - \alpha*1/m*X'*(X'*\Theta - y) + \lambda/m*(\Theta)$

Here,

α is the learning rate of the Gradient descent optimization function,

λ is the regularization parameter to avoid bias and variance.

Let S be the system which can be implemented as:

S = {s, e, I, O, Fme, Fs, MS, SW, URL, feature_vector, DD, NDD, | Øs}

Where,

- s – Start state (movie database)
- e – End state (rating for movie)
- I – set of input
  - I = {i1, i2}
  - Where,
    - i1 - training set
    - i2 – test set
- O – {[0-10]*}
- Fme – main function
- Fs – Set of friend functions
  - Fs = {f1, f2, f3, f4, f5}
  - Where,
    - f1 – function for accepting a movie data as input
    - f2 – function for pre-processing of movie database(feature normalization)
    - f3 – function for creating the input vector
      - feature_vector = {x1, x2, x3, x4, x5, x6, x7}
      - where,
      - x1 = Producer
      - x2 = Director
      - x3 = Actor
      - x4 = Actress
      - x5 = Genre
      - x6 = Runtime

Volume 3 Issue IV, April 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

x7 = Month of release

f4 – function to predict rating
f5 – function for displaying the rating and classifying as hit or flop

IV. DESIGN AND ANALYSIS OF SYSTEM

TABLE I
DATASET USED

Dataset	Examples
Training Set	500
Cross Validation Set	500

The overall architecture of the system can be summed up as shown in the figure below:

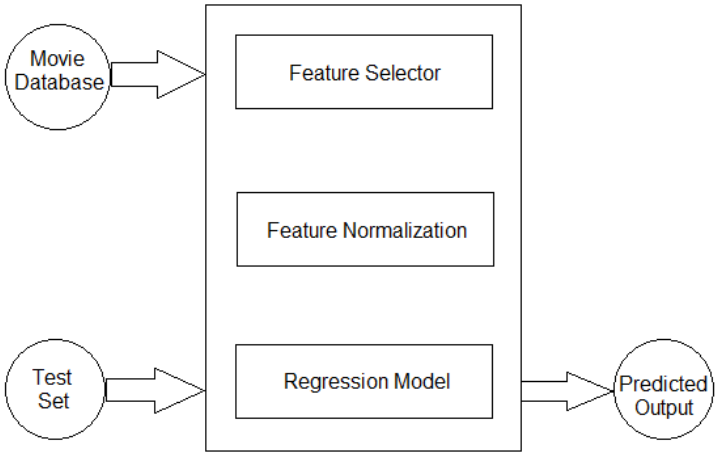


Fig1. Model Architecture

The entire system predicts the future ratings based on the training set of 500 movie details. This movie set was pruned to select a set of features that have been found to make a major impact on the success or failure of a film. After the identification of seven such features all the producers, directors, actors and actresses were rated based on their past performance at the Box Office. Similarly months of the year and runtimes of movie were assigned a score based on the same criteria. After obtaining a numerical equivalent of the movie database, the features were normalized to value in a range of -1 to 1

Feature normalize,

$x^j_i = (x^j_i - mean^j)/range^j$

where,

$mean^j$ is the average of x^j feature of all movies

$range^j$ is the range (largest value – smallest value) of x^j feature of all movies

Such normalized features were feed to the linear regression system.

SYSTEM FLOW (Algorithm):

Input: Movie database, User input film with feature values

Output: Rating of user entered film

1. Feature selection

2. Feature normalization

3. Choose a regression model that fits best to the training set. I chose Multivariate form of Linear Regression

4. Compute Jcost (**Θ**) function and keep minimizing it using Gradient Descent.

5. In step 4 different values of learning rate and lambda were experimented and alpha = 0.01 and lambda= 280 were fixed as these value gave the best result.

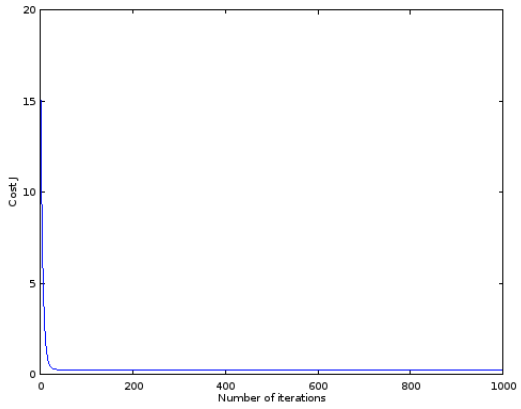
6. Error factor ε was observed to be -6 and hence a value of -6 was subtracted from the hypothesis function for accurate prediction.

7. System was tested with a cross validation set of 500 examples and a accuracy of 70.4% was achieved.

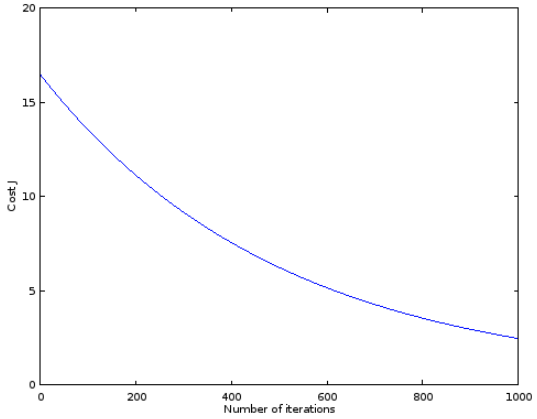
V. DISCUSSION ON IMPLEMENTATION RESULTS

Gradient Descent was iterated for 1000 times to set the values of parameters theta with different values of alpha. The results obtained are shown graphically below.

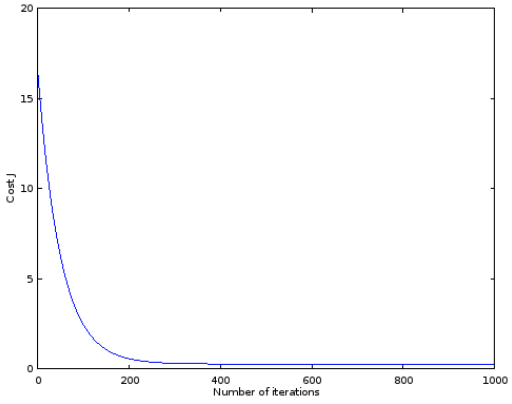
α=0.1



α=0.001



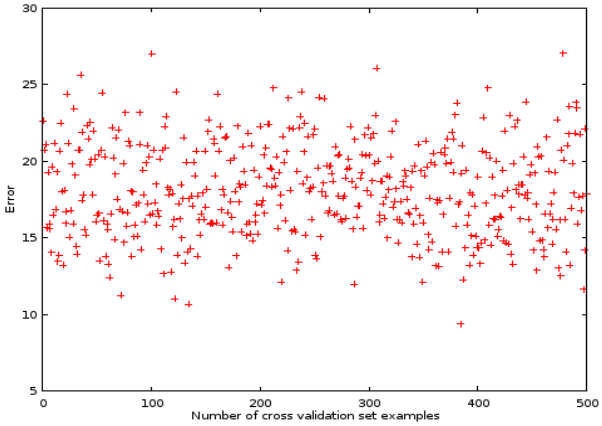
α=0.01



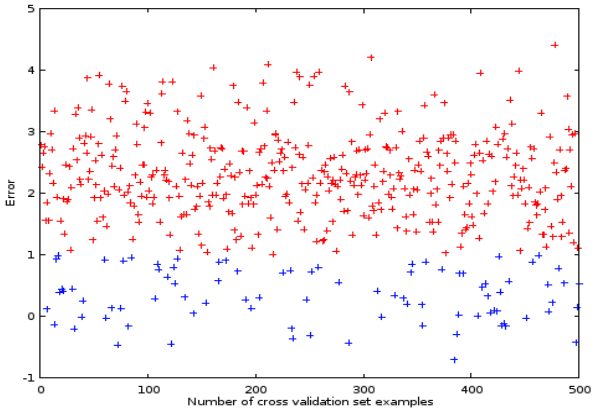
As we can see α =0.01 gives the best optimization graph at around 400 iterations and hence a value of α =0.01 was chose.

Next, different values of λ were tried and following results were obtained.

Volume 3 Issue IV, April 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

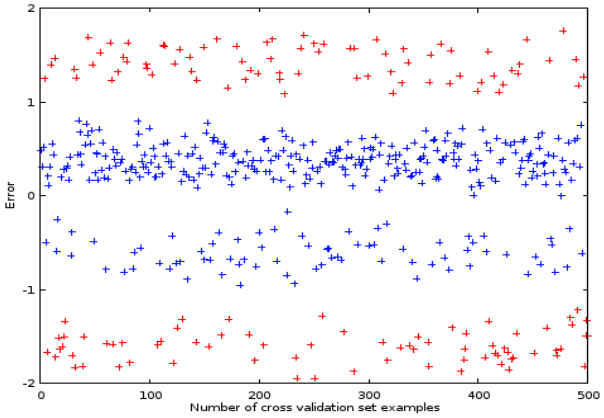


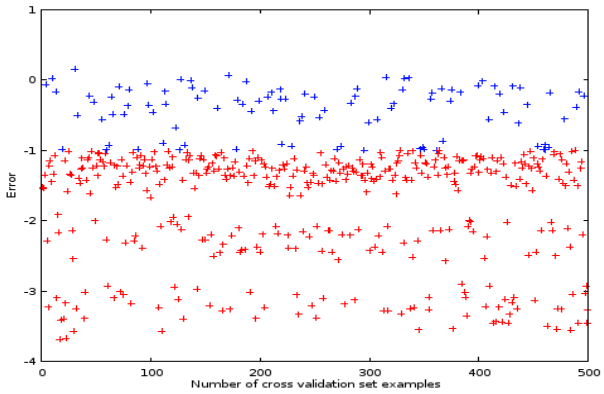
λ=0 (no regularization), error~20



λ =200, error~3

λ=280, |error|<1





λ =400, error~2

Thus, it can be observed from the graphs that λ of 280 gives the most accurate result on cross validation set with average rating error smaller than 1. Hence λ =280 was fixed.

The overall accuracy of the system was found to be 70.4%.

VI. CONCLUSIONS

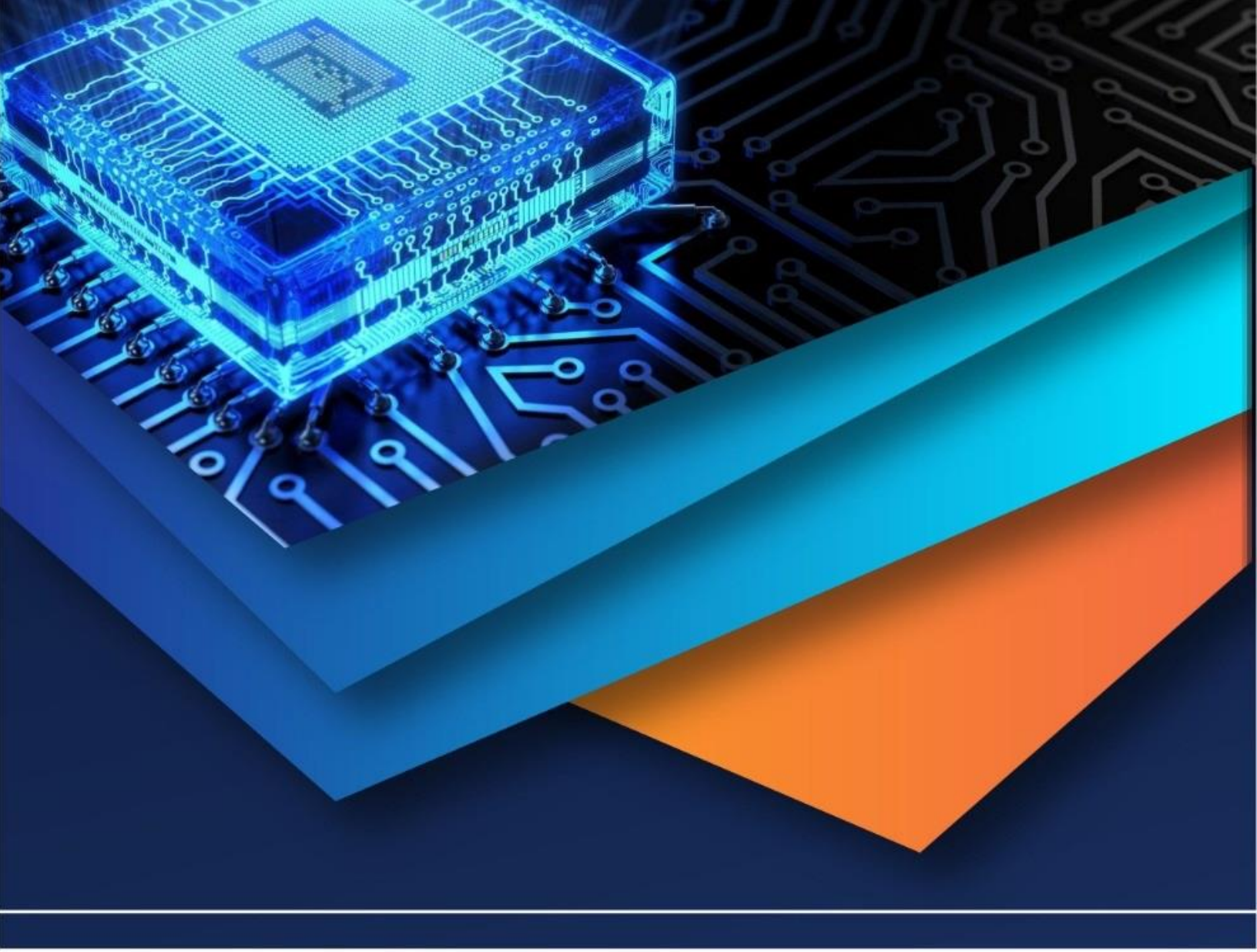
Many Machine Learning Algorithms like Linear Regression can be used to predict various outcomes. Values of α and λ play a vital role in Linear Regression. The system was implemented with 70.4% accuracy. Hence there is still a lot of scope in improving the accuracy of such Algorithms.

VII. ACKNOWLEDGMENT

As students of Stanford's CS-229 MOOC, we would like to thank Coursera for providing such quality material online and our project guide Prof. S. S. Sonawane for her valuable inputs.

REFERENCES

[1] Business Intelligence from Social Media: A Study from the VAST Box Office Challenge,Computer Graphics and Applications, IEEE Oct 2014 (Volume:34 ,Issue: 5),10.1109/MCG.2014.61
[2] Predicting iPhone Sales from iPhone Tweets,2014 IEEE 18th International Enterprise Distributed Object Computing Conference, Pages 8190, 10.1109/EDOC.2014.20
[3] Content Based Prediction of Movie Style, Aesthetics, and Affect: Data Set and Baseline Experiments, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 16, NO. 8, DECEMBER 2014
[4] Improving motion vector prediction using linear regression, Proceedings of the 5th International Symposium on Communications, 10.1109/ISCCSP.2012.6217750,Control and Signal Processing IEEE, ISCCSP 2012, Rome, Italy, 24 May 2012



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY