Mid Term Report

On

PREDICTING MOVIE SUCCESS BASED ON IMDB DATA

By

Jinesh G	B100476CS
Nithin VR	B100121CS
Pranav M	B100778CS
Sarath Babu P B	B100238CS

Under the guidance of

Ms. Lijiya A

Computer Science And Engineering
NATIONAL INSTITUTE OF TECHNOLOGY CALICUT
Calicut, Kerala 673601

Winter Semester 2014

Abstract

American film studios collectively generate several hundred movies every year, making the United States the third most prolific producer of films in the world .And the budget of these movie is of the order of hundreds of millions of dollars, thus their box office success is absolutely essential for the survival of the industry. Knowing which movies are likely to succeed and which are likely to fail beforehand the release could benefit the production houses greatly because it will enable them to focus their advertising campaigns which itself cost millions of dollars accordingly. And it could also help them to know when it is most appropriate to release a movie by looking at the overall market. So the prediction of movie success is of great importance to the industry.

Machine learning algorithms are widely used to make predictions such as growth in the stock market, demand for products, nature of tumors etc. The goal of this project is to use data available in the Internet to predict the revenue of the movies and the critical success as well. Critical success will be represented by the IMDB ratings. The main challenge is to find the appropriate features to be considered for making the predictions. This problem is modelled as a linear regression problem. A supervised learning approach is likely to be followed.

PROBLEM STATEMENT

The aim of the project is to predict the box office revenue and critical success of new Hollywood movies using machine learning techniques. The IMDB rating after at least one month after the release is considered as the index of critical success. We intend to make two predictions one before the release and the second after one week of release.

Contents

Chapter 1

INTRODUCTION

In the United States of America 1000s of films are released every year.Since the 1920s, the American film industry has grossed more money every year than that of any other country. Cinema in America is a multi-billion dollar industry where even single films earn over a billion dollars. Large production houses control most of the film industry, with billions of dollars spent on advertisements alone. Advertisements campaigns contribute heavily to the total budget of the movies. And sometimes the investment result in heavy losses to the producers. Warner Brothers one of the largest production houses had a fall in their revenues last year despite the inflation and increased number of movies released. If it was somehow possible to know beforehand the likelihood of success of the movies the production houses could inturn 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 its not. This shows a dire need for such a software to be developed. Many have tried a various methods to accomplish this goal of predicting the movie revenues. Techniques such as social media sentiment analysis have been used in the past. None of the studies thus far has succeeded in suggesting a model good enough to be used in practice. In this project we attempt to use IMDB data to predict the gross revenue of the movies as well as their IMDB rating.

Chapter 2

LITERATURE SURVEY

Lot of work has been done in the area and too many people have attempted to use statistics in combination with machine learning algorithms to predict various entities.

- Opinion Finder, a project conducted at University of Pittsburg [5] attempted to use tweets to predict market sentiments. Their approach was to give different weightage to the adjectives in the tweets. They assigned values ranging from +2 to -2 to certain words and took the average weightage to make predictions.
- In their paper presented in the 49th annual meeting for ACN in 2011 [6]. Andrew Ng, Christopher Potts and their colleagues pointed out the importance of using a mix of supervised and unsupervised approaches to learn word vectors capturing semantic termdocument information as well as rich sentiment content.
- Independant works conducted by Matts Forsell and Darin Im [3] attempted to predict Movie Revenues and had different success rates. Forsells prediction used a slightly more extensive training than Ims. But the prediction success was lower. Darin Im used a smaller training set, however used data that was from very recent movies. Forsell noted that error rates were significantly higher in predicting success of movies with budget lower than 100 million.
- Project conducted by Deniz Demir and Olga Kapralova at stanford [7] used google trends statistics to improve movie gross predictoin. They

tried two different approaches, in one they used google trends statistics in combination with google adwords in the other the used google trends statistics only.Their data set consisted of 800 movies only, this was mailnly because google trends statistics from 2004 onwards only is available. They used Logistic Regression, SVM and Multilayer perceptron approaches. All three approaches yielded average success rate close to 50 percentage, however SVM method correctly predicted 98% of the low rated movies but its success rate for high rated movies(ratings above 6) was a meagre 11%.

- Wenbin Zhang at Sony Brook University,NY used news analysis [8] alone to predict movie revenues without considering any otther parameters including IMDB data. In their approach they tried to find correlations between various paramenters and movoie revenues. They found that prediction models using merely news data can achieve similar performance with models using IMDB data, especially for high-grossing movies. They concluded that adding news analysis features can improve performance of predictions using IMDB data alone.

Chapter 3

DESIGN

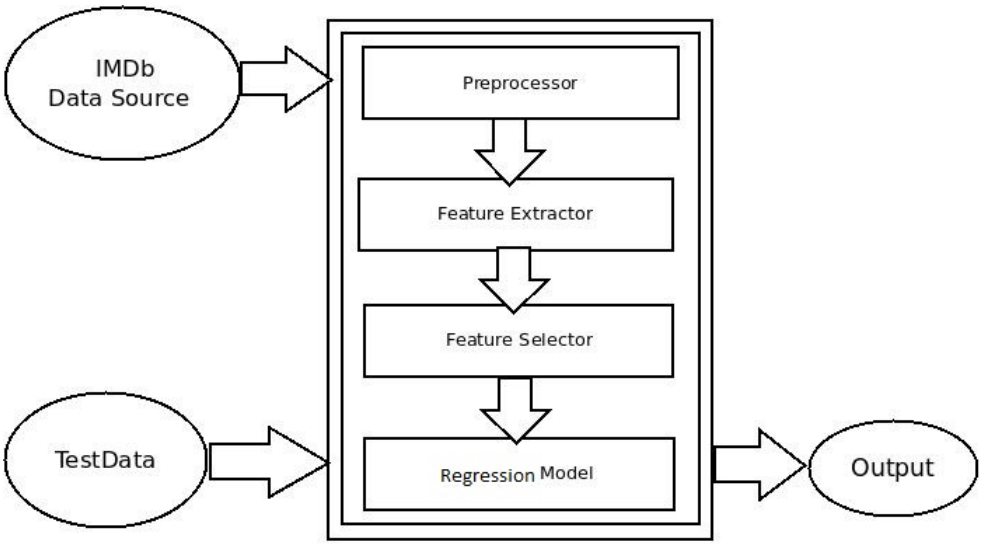

Figure 3.1: General Design

3.1 Introduction

The project can be mainly divided into two stages

- Dataset Collection and Preprocessing.

- Construction of Regression Model.

3.2 Dataset Collection and Preprocessing

Dataset Collection is an important step in all machine learning problems.Preprocessing includes feature extraction and selection.

3.2.1 Dataset Collection

The initial dataset to be used will be collected from IMDb,movies that were released from 2000 to 2012. Among these movies,we only selected the ones that were released in United States and are in English, in the anticipation that we would be able to make more accurate predictions on these movies given that their reviews would also be in English.Then removed the movies which don't have any information about Boxoffice deatils.

3.2.2 Preprocessing

In preprocessing step we intent to eliminate training samples that have insufficient data. Also we intent to ensure that all the 19 genres in the IMDB database are well represented. We also want to filter the data in order to remove redundant and unnecessary information.

3.2.3 Feature Selection

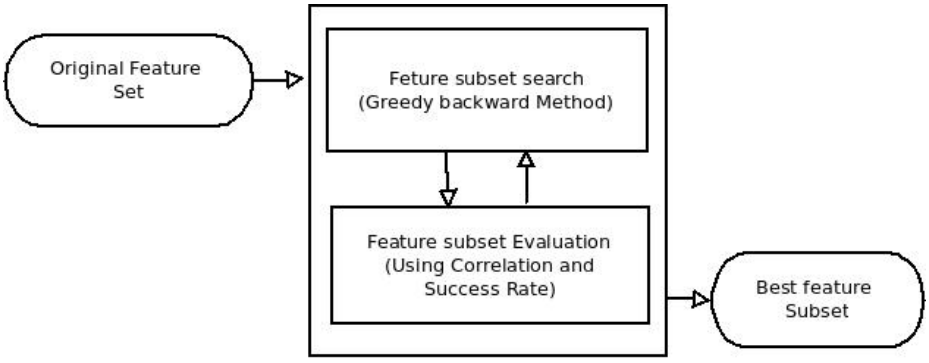

Figure 3.2: Feature Selection

Here we look for correlation between all the different features under consideration and look for correlations with the target variable, ie movie revenue in dollars. We also look for correlation between the features themselves in order to avoid redundancy and irrelevent attributes. Also the extent of correlation is noted.Here we are using **greedy backward procedure** to get best feature subset. The procedure starts with the full set of attributes.At each step, it removes the worst attribute remaining in the set At the end of the process, features that are not correlated to the target variable and those which are redundant are eliminated.

3.3 Construction of Regression Model

Supervised learning technique is adopted for this project. We are using three Regression models to predict the revenue and we will compare the perfomance of the different methods.

We chose Logistic regression as our second method mainly because it generates a multi-class model with linear weights, most directly comparable to the feature weights given by linear regression. To define our classes we draw a histogram of movie revenues to create different buckets for prediction. These buckets are continuous ranges of movie revenues which covers the entire sample space.

3.3.1 Linear Regression Model

In our first model, we use a standard least-squares linear regression. To do this, we intend to use stochastic gradient descent. Once we have trained a set of feature weights, we could then generate gross revenue predictions as follows: $\text{Gross} = \theta_0 + \theta_1 * \text{F}_1 + \theta_2 * \text{F}_2 + ... + \theta_n * \text{F}_n$
where θ_i are the weights, F_i are the features, and n is the number of features.

3.3.2 Logistic Regression model

We chose Logistic regression as our second method mainly because it generates a multi-class model with linear weights, most directly comparable to the feature weights given by linear regression. To define our classes we draw a histogram of movie revenues to create different buckets for prediction. These

buckets are continuous ranges of movie revenues which covers the entire sample space.

3.3.3 Support Vector Machine Regression Model

SVMs can also be applied to regression. SVM Regression try to find a function f(x)that has at most ε deviation from the actually obtained targets y for all the training data, and at the same time is as flat as possibles.SVM Regression do not care about errors as long as they are less than ε .We used linear kernel function to map the data into a high dimensional feature space where linear regression is performed.

Chapter 4

WORK DONE

4.1 Obtaining The Data Set

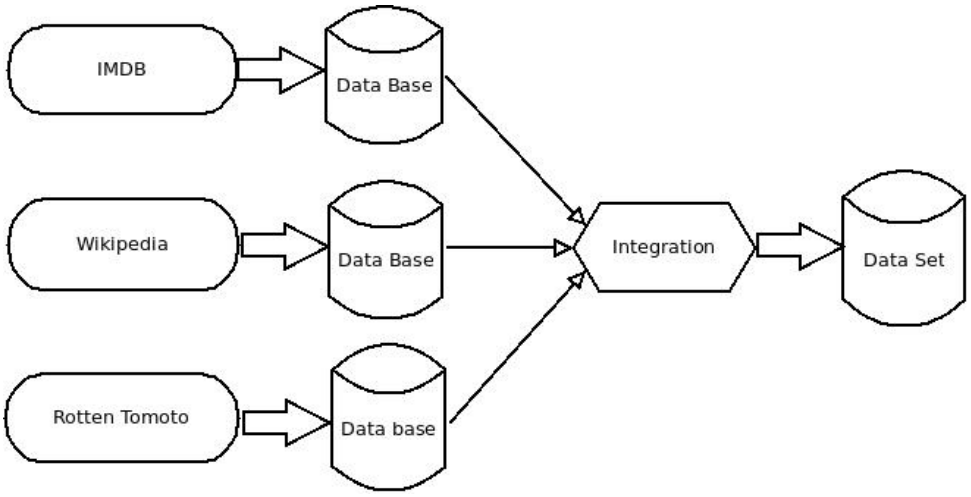

Figure 4.1: Feature Selection

The data for a subset of movies,actors,reviews is provided by imdb via an alternative interface. Python scripts could be used to obtain data from IMDB. The data regarding each movie is spread over two web pages. The main page(imdb.com/title/movieid.html) and box office page. The data is also available at other websites like OMDB and Rotten Tomatoes in JSON format. Eventually we faced a problem with Budget Data. Budget is not

provided by both IMDB and Rotten Tomatoes. So we used wikipedia to obtain Budget details.

4.2 Data Preprocessing

The data we obtained are highly suseptible to noisy, missing and inconsistent data due to huge size and their likely origin from multiple, heterogenous sources[9]. We mainly used IMDB and Rotten Tomatoes and Wikipedia. The main problem with datasets were missing fields. To overcome this missing field problem we adopted a method which uses a measure of central tendancy for the attribute. We used both mean and median as central tendancy. Then removed duplicate entries.

4.2.1 Data Integration

Data obtained from three different resources IMDB, Wikipedia and Rotten Tomatoes was then integrated into one database.

4.2.2 Data Transformation

In this step integrated data are transformed or consolidated so that the regression process may be more efficient and easier. Dataset is mixed with both nominal and numeric attributes but for regression process we need all attributes to be numerical. We used a measure of central tendancy of Box-office revenue to convert corresponding nominal attributes to numerical.

4.2.3 Selecting feature Subset

from the whole dataset we obtained 20 features Implemeted greedy backward procedure to reduce the dataset in to best feature subset. Got a 7 feature subset with maximum output.

4.2.4 Measuring Sentiment Score From Reviews

Implemented a module to compute sentiment score using a technique similar to the one used in the opinion finder project. The method involves assigning weights to commonly used adjectives to obtain an average weightage.

4.3 MACHINE LEARNING

4.3.1 Building a Linear Regression Model

Using the method of Normal equation we built a linear regression model. Used data regarding 950 films for trainig purpose and 100 for testing.

4.3.2 Logistic Regression model

4.3.3 SVM with Linear Kernal

4.4 WEB APPLICATION

4.4.1 Google Appengine

4.4.2 OMBD API

4.4.3 Wikipedia

Chapter 5

RESULT

Chapter 6

CONCLUSION

Bibliography

[1] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification

[2] Sagar V. Mehta, Rose Marie Philip, Aju Thalappillil Scaria, Predicting Movie Rating based on Text Reviews. University of Stanford CS229 Projects, 2011

[3] Darin Im, Minh Thao, Dang Nguyen, Predicting Movie Success in the U.S. market. University of Stanford CS229 Projects, 2011

[4] Suhaas Prasad, Using Social Networks to improve Movie Ratings predictions. University of Stanford CS229 Projects, 2010

[5] Opinion Finder Project, University of Pittsburg. http://mpqa.cs.pitt.edu/opinionfinder/

[6] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts ,Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 142−150

[7] Deniz Demir, Olga Kapralova, Hongze Lai, Predicting IMDB Movie Ratings Using Google Trends.

[8] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis.
Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference,2009

[9] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques 3/e