Movie Gross Revenue Estimation

7th Semester Project

Department of Information Technolgy
Indian Institute of Information Technology
Allahabad, Uttar Pradesh(U.P)

Project Mentor and Guide
Dr. Sonali Agarwal

September 19, 2016

THE TEAM

1. Abhishek Kumar Agrawal(IIT2013128)
2. Shubham Bhendarkar (IIT2013172)
3. Shaiwal Sachdev (IIT2013196)

List of Figures

4.1 Scraping From IMDB Page 6
4.2 Revenue Data From Business Page 6

6.1 Block Diagram of project 8

7.1 Currency Conversions and Dealing with unicode characters 10
7.2 Scraping From Rotten Tomatoes Page 11
7.3 Currency Conversions and Dealing with unicode characters 11

8.1 Formula of MAPE 12

List of Tables

4.1 About the Data 5

7.1 Year vs Movie Ticket Prices 10

8.1 Correlation (Gross Revenue vs Others) 12

10.1 Genre vs Best Combination 15

Contents

1 Introduction **1**

2 Problem Definition **2**

3 Literature Survey **3**

4 Dataset **5**
4.1 Data 5
4.2 Field Information 5

5 Technologies **7**
5.1 Tools Used 7
5.2 Languages Used 7

6 Proposed Approach **8**
6.1 Phases 8
6.2 Block diagram 8

7 Phases of Project **9**
7.1 Data Collection 9
7.2 Data Cleaning 9
7.3 Data Transformation 9

8 Estimation Using Regression Models **12**
8.1 Error Measurement (MAPE) 12

9 Genre Based Regression Analysis **13**

10 Result **15**

11 Conclusion **16**

12 Things to Do After Mid Sem **17**

Chapter 1

Introduction

Film industry is a big business in United States.It is one of the biggest player in the entertainment industry.Predicting the gross revenue of a movie beforehand is required by a lot of people.Till now there is no computational model that can effectively predict the Gross revenue movie will be able to collect.This depends on a lot of factors like release date,number of available theatre screens,budget of the film,cast, MPAA rating ,release year,etc.

We can develop a model by using the history revenue data for previous year movies and predict the gross revenue, movie will be able to get by using different set of features.

This model is useful for :

1. Movie producers and Production studios as by looking at estimated values of revenue, they can take different steps on deciding the budget for things like marketing,promotion,etc.

2. Movie theatres as they can also estimate the amount of money they will be able to collect on screening the movie.

If the estimated revenue is very low , studios may increase their promotion budget and may even think of releasing the movie outside the domestic space. Studios try to release it in more theatres , they display banners , Actors promote the movie in TV Shows, sometimes they also do publicity stunt. Overall, every movie maker wants to earn more and more money.

In this project we will collect the revenue data and other details about previous year movies and make a model that will be able to predict the gross revenue of the film.

Chapter 2

Problem Definition

The goal of this project is to develop a model that will be able to estimate the Box Office Gross Revenue of a film after its first weekend of release.

The analysis is based on USA region only.

Chapter 3

Literature Survey

A Lot of factors that affect the revenue prediction have been studied by different researchers.

Do Critical Reviews Really Matter? As mentioned in [7] even though some reviewers may have different tastes than the people reading their reviews, consumers still read the reviews before going to watch the movie to know whether it is worth their money or not. This tells us that critic ratings do matter in the revenue prediction model.

Forswell in [9] collected public information available on IMDB site for movies after January 1st, 1990 and only included those having total box office revenue greater than 100,000 dollars totaling 2500 movies.He then used linear regression model using features first week end revenue,budget and number of available theatre screens.It did not give very good result and when he tried splitting the dataset, number of training samples decreased and performance was reduced.

Robert in [11] divided the set of features in simple , complex and sentiment where simple is numeric only, complex is numeric and text based and sentiment includes all.He thought that this problem can be modeled both as regression and classification. He used logistic regression for classification by making classes by dividing the range of min and max value of gross revenue into buckets.After doing more analysis he found that text and sentiment features were insufficient to predict the revenue.

Some researchers like in [3] tried to predict the revenue before release or after 1st week of revenue. They analysed different features like production budget of the film,revenue of the 1st weekend , sequel movie or not,star power,MPAA(Motion Picture Association of America film rating system) rating etc. He analysed the change in gross revenue with respect to rating of the movie.He tried to give weights to these categorical variables and used linear regression to do the prediction.

Vitelli in [14] tried to create a set of features and did extract values from graphical properties of the actor-actor, actor-movie, and movie-movie relationships.

Some tried to predict the revenue before the release and some after 1st week of release.Budget of the film seems to different effects on different genre movies. Mostly big budget films which are Action,Animation, Adventure tend to attract a large amount of audience whereas on the other hand movies which even after spending a lot of money like which are biography,drama,etc tend to earn less.This fact has been supported in [5].

Genre and revenue analysis was done. In Anast [1] who tried to showcase some relationship between genre and revenue. Prag and Casavant [2] showed a negative relationship for drama and revenue.

Pre release revenue Prediction although looks attractive but it is a very difficult

job.Extracting information from a variety of sources like blogs,youtube trailers comments,estimating the people's response before the movie is released.Use the number of viewers and users who edit the page in wikipedia was used by Marton in [8] to find or estimate the popularity parameter in his model to find the pre release revenue. To find the estimate after 1st week gives us more predictability as it is almost twenty to twenty five percent of the gross domestic revenue. It gives us insight that we will get more accurate or better results if we estimate it after 1st week of release or Post Release.

Dursun Delen in [10] used Neural Network with features like MPAA rating,genre,star value,sequel,special value but for prediction of pre-release revenue. It was used because neural networks can handle a mix of continuous and discrete values pretty well.

Some people also tried to predict the pre release box office revenue using movie scripts as in [6] and also tried using the tweets data as in [15], blog data to anaylse the behavior of the people but these methods did not give satisfactory results.Thorsten in [12] also anaylses the negative and postive effect of tweets on the behavior of people which in turn affects the box office revenue.

Till now everyone was concentrating on Domestic revenue only but researchers as in [4] tried to estimate the Foreign Revenue the movie will be able to collect. This depends on a lot of factors like domestic success of the movie , language adaptability, cultural differences,MPAA rating differences between domestic and foreign country.

Till now all the researchers were considering all the different factors that affect the movie success.But Thorsten in [13] did a study on inter dependence between these features whether success or effect of one feature can affect the effect of other on revenue.For example,does advertising influence the box office revenue directly by creating a media presence, or indirectly through impacting consumers quality perceptions of the movie?

Overall all the researchers tried to estimate the Gross Box office revenue movie will be able to collect.Some considered it to be regression problem and few as classification problem.None of them tried to analyse it genre specific. As if saying that all these features like MPAA rating , production budget of the film, marketing budget of the film,critic rating in the first week are not all important for all genre. Biography or Drama movie may be good spending a lot of money but would not be able to generate a lot of revenue.Likewise, only action and animation big Budget films attract a lot of audience not true for other. Some superhero movies like Super Man and others saying Big Action Movies people don't care about critic ratings they just see it. We should not consider Critic Rating in estimating Action movie revenue. And , budget of the film should not a parameter in Drama or Biography films. Talking about MPAA ratings, movies that are PG-13 , R and action,adventure,horror affect the revenue. They should be considered as a parameter. Whereas movies made for children like G rated , they when released in US mostly collect almost equal amount of revenue.Here we are focusing on post release domestic revenue prediction only.

Chapter 4

Dataset

Initially we did not had the dataset so we crawled and scraped the data about different movies from different websites and mergea all the information collected about each movie into one.

4.1 Data

Name	Information
Total Movies Crawled	3000
Websites Used	IMDB,Rotten Tomatoes ,BoxOfficeMojo, ecosia
Genre	Action,Animation,Adventure,Horror,Sci-Fi,Comedy,Music,Documentary
Business Data	Opening Weekend Revenue, Budget, Gross Total Revenue

Table 4.1: About the Data

All the data collected is then merged into one to make a database containing all the information about the movies. We also scraped the business page for each movie and collected revenue information. We are focusing on USA region only.And the gross revenue is also domestic total gross revenue for USA region only.

We collected the data for movies released 2000 to 2015 only.

4.2 Field Information

The fields collected are :

1. Title of Movie
2. Genre (Action,Animation,Adventure,Horror,etc.)
3. Release Date
4. Budget of the movie
5. Opening Weekend revenue for USA

6. Number of Screens in opening weekend
7. Total runtime of the movie
8. MPAA film rating
9. Rotten Tomato Critic Rating
10. Userrating from Rotten Tomato
11. Gross Domestic Revenue for USA
12. MetaScore from Metacritic
13. Popularity of the Movie
14. IMDB User rating

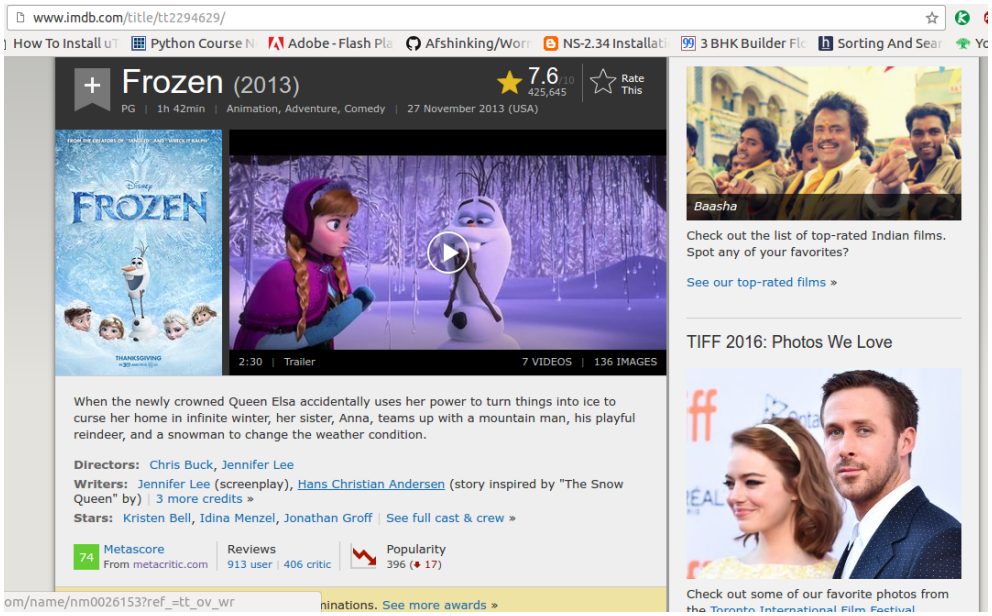

Figure 4.1: Scraping From IMDB Page

Figure 4.2: Revenue Data From Business Page

Chapter 5

Technologies

5.1 Tools Used

1. Scikit Learn
2. LaTeX
3. miniconda
4. matplotLib
5. BeautifulSoup
6. Google API
7. Rotten Tomatoes API

5.2 Languages Used

1. Python

Chapter 6

Proposed Approach

6.1 Phases

The project comprises of following four phases

1. Data Collection
2. Data Cleaning(Preprocessing)
3. Data Transformation (Normalization)
4. Estimation Using Regression and Classification Models

6.2 Block diagram

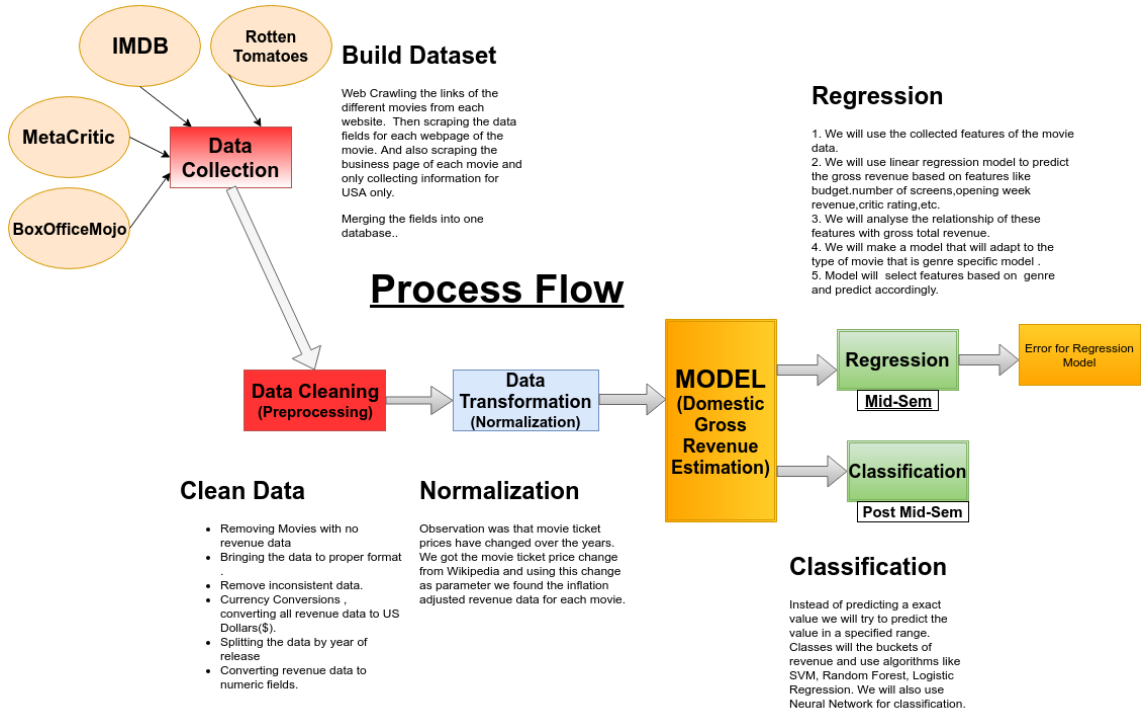

Figure 6.1: Block Diagram of project

Chapter 7

Phases of Project

7.1 Data Collection

We crawled and scraped movie information for about 3000 movies and collected different fields from different websites. We collected the information by year of release about 200 movies per year from 2000 to 2015 .We then dumped all the information in the form of json. We then split the dataset into different subsets by the type or genre of the movie.

7.2 Data Cleaning

1. Removing Movies data with no revenue data.
2. Bringing the data to proper format like removing commas,dollar sign, dealing with words like "million" , that is anything other than numeric data.
3. Dealing with Unicode characters in title of the movie.
4. Removing inconsistent data.
5. Budget of the movie was in different currencies . Currency Conversions had to be done for converting all this data to US Dollars. More than 30 currencies were found in the 3000 movies collected.
6. Splitting the data by year of release and genre.
7. Converting revenue data to numeric fields.

7.3 Data Transformation

We observed that movie ticket prices have changed over the years.Due to inflation price of tickets have changed so much over the years. We got the movie ticket price change from Wikipedia. We used this change in ticket price and considered this as inflation parameter and calculated the inflation adjusted revenue data for each movie.

Ticket prices changed so rapidly with price of 5.39 USD in year 2000 to about 8.66 USD in 2016.

Year	Movie Ticket Price(USD)
2016	8.66
2015	8.43
2014	8.17
2013	8.13
2012	7.96
2011	7.93
2010	7.89
2009	7.50
2008	7.18
2007	6.88
2006	6.55
2005	6.41
2004	6.21
2003	6.03
2002	5.81
2001	5.66
2000	5.39

Table 7.1: Year vs Movie Ticket Prices

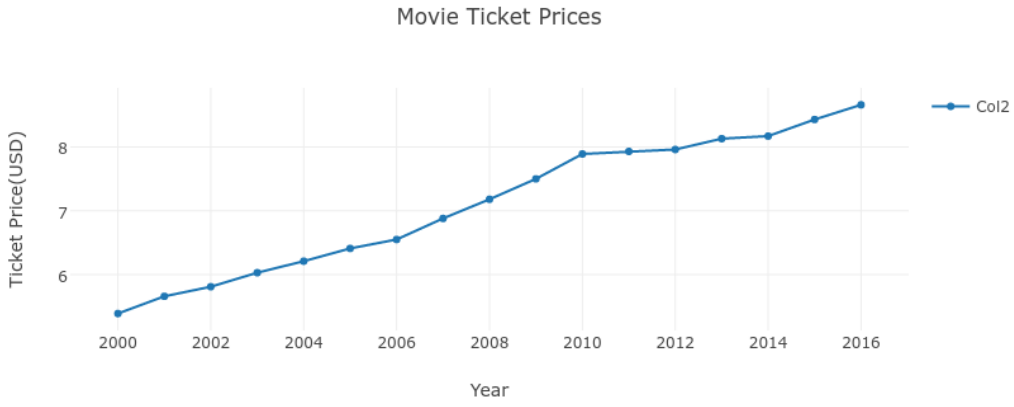

Figure 7.1: Currency Conversions and Dealing with unicode characters

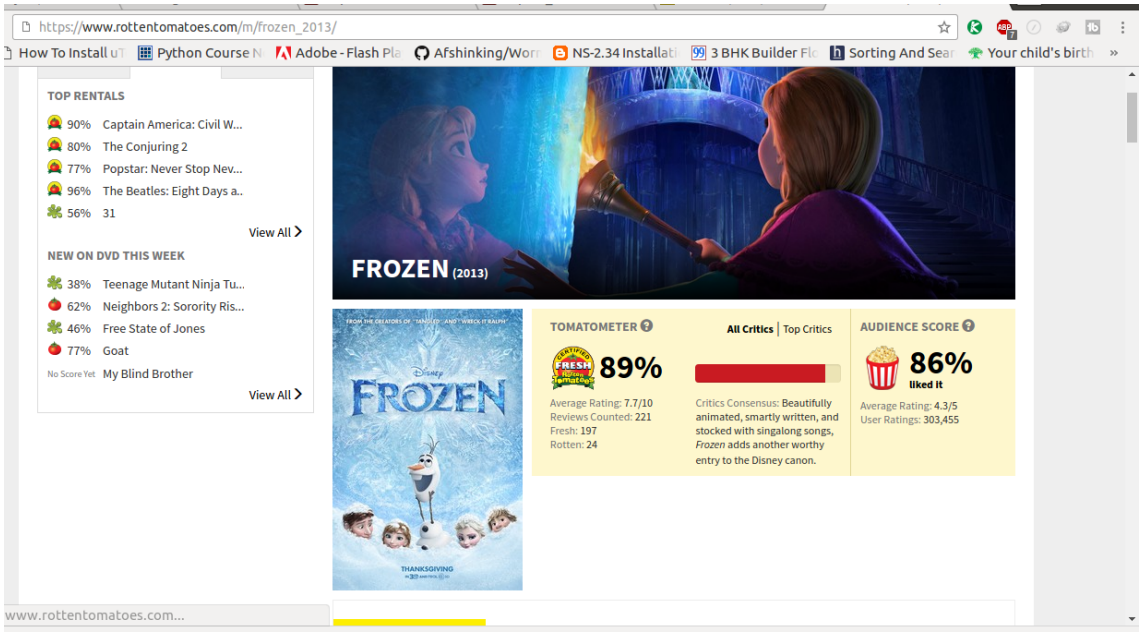

Figure 7.2: Scraping From Rotten Tomatoes Page

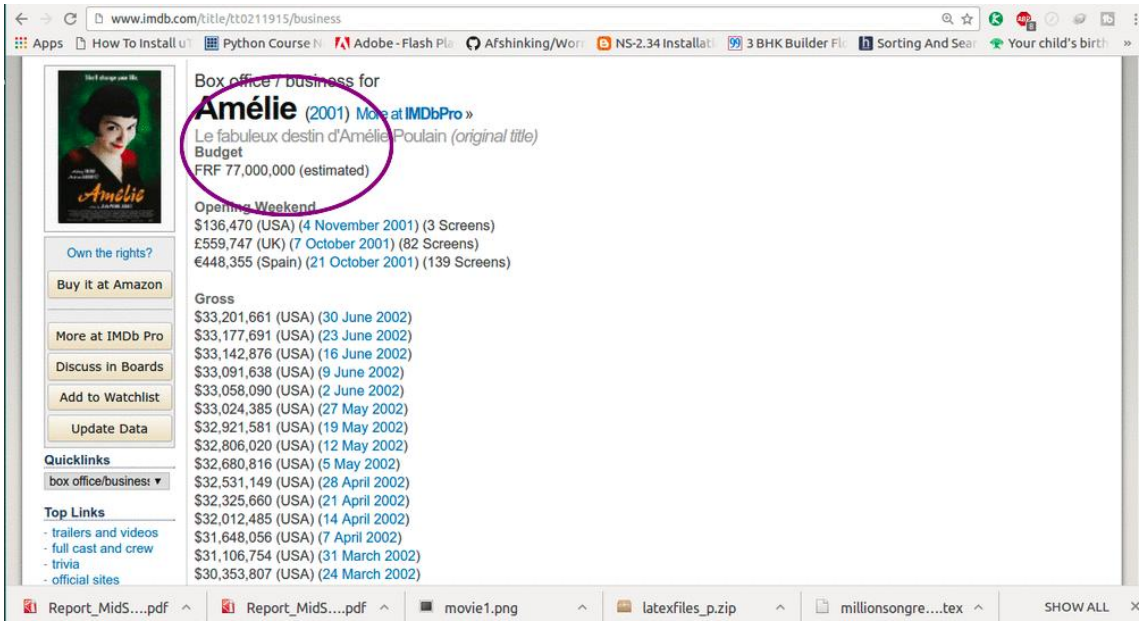

Figure 7.3: Currency Conversions and Dealing with unicode characters

Chapter 8

Estimation Using Regression Models

Assuming linear relationship between gross revenue and other features , we first analyzed the linear dependence of gross revenue on features like budget of the movie, number of screens ,opening weekend revenue, metascore, tomato ratings,user reviews,etc.

We calculated the correlation between gross revenue and other features.

Budget	Opening Weekend	Screens
0.691	0.922	0.553

Table 8.1: Correlation (Gross Revenue vs Others)

8.1 Error Measurement (MAPE)

Mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD) is the mean of the percentage error for each sample.

From the above table, we came to a conclusion that gross revenue largely depends on the opening weekend revenue as high correlation (0.922).Opening weekend revenue actually accounts for about 20 to 25 percent of the total revenue.

$$\mathrm{M} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|,$$

Figure 8.1: Formula of MAPE

We then applied a multiple linear regression model with features budget of the movie, number of screens ,opening weekend revenue. We then calculated the MAPE error between the predicted and real gross values.

We got an error percentage of 130 percent. This is huge error.In order to improve this error, we did genre based regression analysis.

Chapter 9

Genre Based Regression Analysis

According to [2] journal paper , different types of movies respond to different parameters differently.

Like Budget of the movie maybe important for Action or Animation movie but not for a Horror Movie. UserRatings or UserReviews affect movies like documentary or Horror but not Adventure or Comedy movies.

We thought of finding the percentage error for each genre for each combination and try to find the best set of features that describe a genre.

Following things we did:

1. We first split our data set into different parts based on genre basis.
2. We then find out all the subsets of the set of features.
3. We then calculated the error for each genre for each subset of features. We are doing this to find the best combination that will give us minimum error.
4. 10 fold cross validation is used and then we find the mean value of MAPE for all folds.
5. This way we are making a model that will change its set of input parameters required based on type of movie .
6. Calculate error for each type of movie for each combination.
7. Picking up the combination with minimum percentage error for each combination

Genres we considered are :

1. Action
2. Adventure
3. Comedy
4. Animation
5. Crime
6. Horror
7. Documentary
8. Biography
9. Drama
10. Sci-Fi
11. Romance
12. Thriller
13. Family
14. Fantasy
15. History
16. Mystery
17. Sport
18. Music

Chapter 10

Result

After doing genre based analysis,we got the following result for each genre showing the best combination of features for each.

Genre	Best Combination	MAPE
Action	Opening Weekend,Popularity	46.45
Adventure	Budget,Opening Weekend	49.18
Comedy	Budget,Opening Weekend	50.26
Animation	Budget,Opening Weekend	36.693
Crime	Budget,Opening Weekend	80.95
Horror	Opening Weekend,Rotten User Votes	23.47
Documentary	popularity,tomatoRating,IMDB Userrating,Rotten User Votes	42.57
Biography	Budget,Opening Weekend,popularity,tomatoMeter,tomatoRating	95.35
Drama	Budget,Opening Weekend,Metascore	118.02
Sci-Fi	Opening Weekend	24.20
Romance	Budget,Opening Weekend,Rotten User Votes	108.41
Thriller	Budget,Opening Weekend	47.17
Family	Budget,Opening Weekend	39.22
Fantasy	Budget,Opening Weekend,Screens,Popularity,IMDB Rating	47.94
History	Opening Weekend,Popularity,tomatoRating	62.38
Mystery	Budget,Opening Weekend,Popularity,User Ratings	49.89
Sport	Budget ,Opening Weekend,popularity,tomatoMeter,tomato Rating	42.55
Music	Budget,Opening Weekend,tomatoMeter,Rotten User Votes	60.78

Table 10.1: Genre vs Best Combination

Chapter 11

Conclusion

From the result table showing the genre and best combinations with percentage error:

1. Shows the Budget Of Movie don't seem to affect the gross revenue of Horror,Documentary,Sci-Fi,History movies therefore it should not be taken as a parameter in regression while doing estimation.
2. Opening Weekend is the most important feature that is present in all of the genres.
3. Critic Rating don't seem to affect the performance of a lot of movies. Only Documentary,Biography,Drama,History Sport,Music were affected a bit.
4. Budget and Opening Weekend seem to make the best combination.

Chapter 12

Things to Do After Mid Sem

Currently, we have calculated the MAPE error on the genre basis.These things we will do post mid semester:

1. Finding other models of regression trying to decrease the error.
2. We are looking forward to decrease percentage error by collecting more data and removing outliers.
3. Dealing with movies with more than one genre.
4. Making use of categorical variables like MPAA rating,etc.
5. Instead of predicting a exact value we will try to predict the value in a specified range. Classes will the buckets of gross revenue.Basically, saying this can also be seen as a classification problem.

References

[1] *Anast, P. 1967. âDifferential movie appeals as correlates of attendanceâ, Journalism Quarterly.*

[2] *Prag, J.J. Casavant, J. 1994. âAn empirical study of the determinants of revenues and marketing expenditure in the motion picture industryâ, Journal of Cultural Economics,.*

[3] *Simonoff, J. S. and Sparrow, I. R. Predicting movie grosses: Winners and losers, blockbusters and sleepers. In Chance, 2000.*

[4] Ryan Compton Brian de Silva. *Prediction of Foreign Box Office Revenues Based on Wikipedia Page Activity.* 2014.

[5] Luis Cabral and Gabriel Natividad. *Box-Office Demand: The Importance of Being.* 2013.

[6] Z. John Zhang Jehoshua Eliashberg, Sam K. Hui. *Assessing Box Office Performance Using Movie Scripts.* 2014.

[7] Alec Kennedy. *Predicting Success: Do Critical Reviews Really Matter?* 2010.

[8] Janos Kertesz MÃ¡rton Mestyan, Taha Yasseri. *Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data.* 2013.

[9] Anahita Sidhwa Nikhil Apte, Mats Forssell. *Predicting Movie Revenue?* 2011.

[10] Dursun Delen Ramesh Sharda. *Predicting box-office success of motion pictures with neural networks.* 2006.

[11] David Cummings Steven Yoo, Robert Kanter. *Predicting Movie Revenue from IMDb Data.* 2011.

[12] Fabian Feldhaus Thorsten Hennig Thurau, Caroline Wiertz. *Does Twitter matter? The impact of microblogging word of mouth on consumersâ adoption of new movies.* 2014.

[13] Gianfranco Walsh Thorsten Hennig-Thurau, Mark B. Houston. *Determinants of Motion Picture Box Office and Profitability: An Interrelationship Approach.* 2008.

[14] Matt Vitelli. *Predicting Box Office Revenue for Movies.* 2012.

[15] Ross Maciejewski Yafeng Lu, Feng Wang. *Business Intelligence from Social Media:A Study from the VAST Box Office Challenge.* 2014.