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

November 30, 2016

Candidate's Declaration

We hereby declare that the project entitled "Movie Gross Revenue Estimation" submitted by us to Indian Institute of Information Technology ,Allahabad as a 7th Semester Project is a an authenticated record of our original project work carried out of us under the guidance of Dr. Sonali Agarwal. Due references and acknowledgments have been made in the text to all the resources and material used.We assure that this report is free from plagiarism and all the work was done was done in full compliance with the requirements and constraints of the prescribed curriculum.

Date: November 30, 2016

Authors:
Abhishek Kumar Agrawal(IIT2013128)
Shubham Bhendarkar (IIT2013172)
Shaiwal Sachdev (IIT2013196)

Supervisor's Certificate

This is to certify that the Project entitled ”Movie Gross Revenue Estimation” is a bonafide record of research project by Shaiwal (IIT2013196),Abhishek(IIT2013128) and Shubham(IIT2013172) under my supervision and submitted to Indian Institute of Information Technology ,Allahabad as a 7th Semester Project.

Date: November 30, 2016

(Dr. Sonali Agarwal)

Acknowledgement

We have taken efforts in completing this project.We sincerely appreciate the inspiration, support and guidance of all those people who have been instrumental in making this project a success. This has been possible with the co-ordination and working together as a team. Everyone in team has contributed in making this project.I am highly indebted to Dr. Sonali Agarwal for their guidance and constant supervision as well as for providing necessary information regarding the project and also for her support in completing this project.

THE TEAM

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

List of Figures

List of Tables

List of Equations

Contents

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 using the public information available 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 [12] 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 [15] 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 [11] 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 [16], blog data to anaylse the behavior of the people but these methods did not give satisfactory results.Thorsten in [13] 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 [14] 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 merged all the information collected about each movie into one.

4.1 Data

Name	Information
Total Movies Crawled	3000
Websites Used	IMDB,Rotten Tomatoes,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

4.2.1 General Features

1. Title of the Movie
2. Genre of the Movie(Action,Adventure,Comedy,..)
3. Release Date
4. Total Runtime of the Movie
5. Year of Release
6. MPAA Rating(Motion Picture Association of America film)

4.2.2 Important Features

1. Budget Of the Movie(USD)

2. Opening Weekend Revenue(USD)

3. Number Of Screens in Opening Weekend

4. Movie Gross Domestic Revenue(USD)

4.2.3 Critic View

1. Rotten Tomatoes(TomatoMeter,TomatoRating)

2. MetaCritic(MetaScore)

4.2.4 User View

1. From Rotten Tomatoes(UserMeter,UserRating)

2. From IMDB

4.2.5 Popularity

1. From IMDB

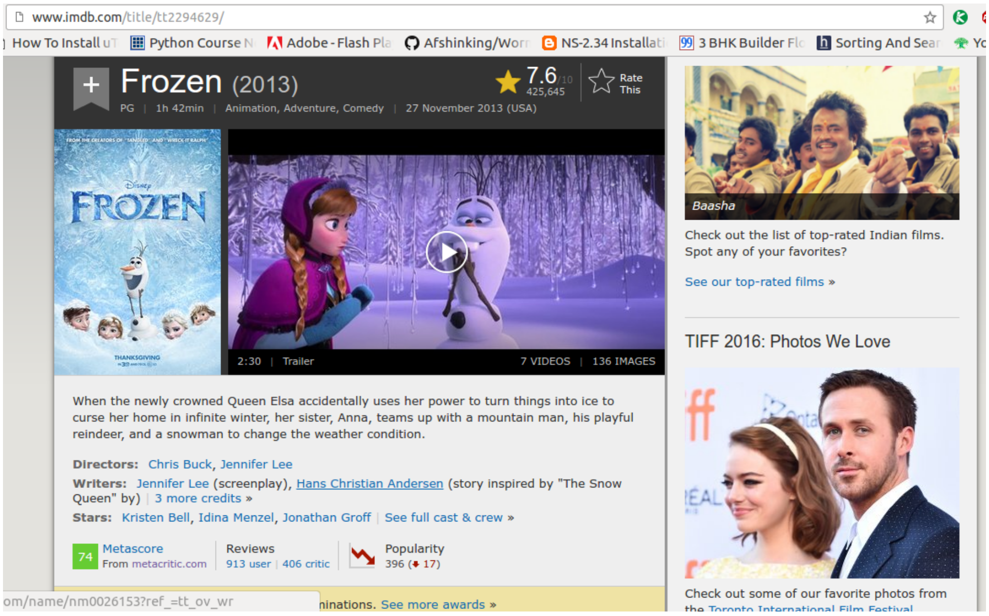
2. From Rotten Tomatoes

4.3 Fields Description

Going through the **Critic View**, we come across TomatoMeter and TomatoRating. TomatoMeter is based on the on the published opinions of hundreds of film critics and is a trusted measurement of movie quality for millions of moviegoers.Meter represents the percentage of critic reviews that were positive that is they rated is above 5 out of 10.TomatorRating is the average critic rating on a ten point scale.MetaScore is a critic score calculated by MetaCritic. com based on critics reviews.Usually the movies called as very good movies have scores above 70.

Going through the **User View**, UserMeter is percentage of users who rated it positive. UserRating is the average user rating on a five point scale.IMDB user rating is done by users or viewers on a ten point scale.

Going through the **Popularity** section, Popularity score from Rotten Tomatoes gives us the number of users who wanted to watch the movie and those who watched and rated it.



Frozen (2013)

7.6/10 425,645

PG | 1h 42min | Animation, Adventure, Comedy | 27 November 2013 (USA)

When the newly crowned Queen Elsa accidentally uses her power to turn things into ice to curse her home in infinite winter, her sister, Anna, teams up with a mountain man, his playful reindeer, and a snowman to change the weather condition.

Directors: Chris Buck, Jennifer Lee
Writers: Jennifer Lee (screenplay), Hans Christian Andersen (story inspired by "The Snow Queen" by) | 3 more credits »
Stars: Kristen Bell, Idina Menzel, Jonathan Groff | See full cast & crew »

74 Metascore From metacritic.com | Reviews 913 user | 406 critic | Popularity 396 (↓ 17)

ninations. See more awards »

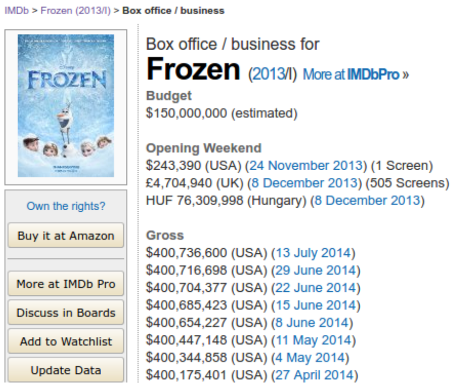
Check out the list of top-rated Indian films. Spot any of your favorites?

See our top-rated films »

TIFF 2016: Photos We Love

Check out some of our favorite photos from the Toronto International Film Festival

Figure 4.1: Scraping From IMDB Page (Source:IMDB)



IMDb > Frozen (2013/I) > **Box office / business**

Box office / business for
Frozen (2013/I) More at IMDbPro »

Budget
$150,000,000 (estimated)

Opening Weekend
$243,390 (USA) (24 November 2013) (1 Screen)
£4,704,940 (UK) (8 December 2013) (505 Screens)
HUF 76,309,998 (Hungary) (8 December 2013)

Gross
$400,736,600 (USA) (13 July 2014)
$400,716,698 (USA) (29 June 2014)
$400,704,377 (USA) (22 June 2014)
$400,685,423 (USA) (15 June 2014)
$400,654,227 (USA) (8 June 2014)
$400,447,148 (USA) (11 May 2014)
$400,344,858 (USA) (4 May 2014)
$400,175,401 (USA) (27 April 2014)

Figure 4.2: Revenue Data From Business Page (Source:IMDB)

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
2. Java

5.3 Websites Used to Collect Data

1. `http://www.imdb.com/`
2. `https://www.rottentomatoes.com/`
3. `https://www.ecosia.org/`

Chapter 6

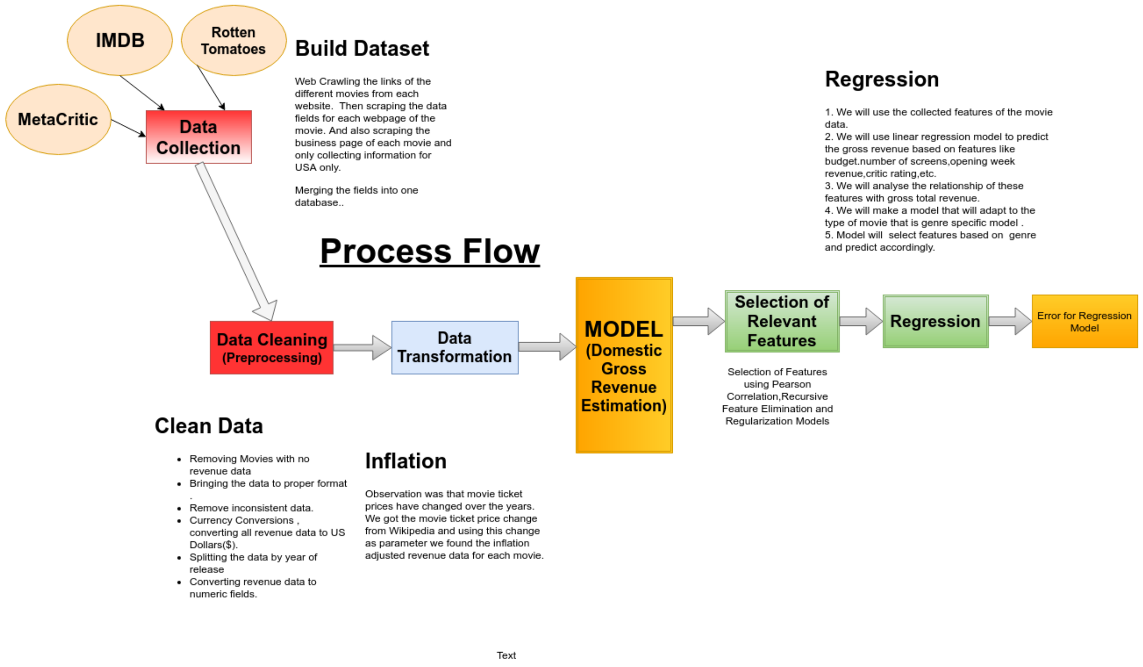
Proposed Approach

6.1 Phases

The project comprises of following four phases

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

6.2 Block diagram


IMDB
Rotten Tomatoes
MetaCritic
Data Collection
Process Flow
Data Cleaning (Preprocessing)
Data Transformation
MODEL (Domestic Gross Revenue Estimation)
Selection of Relevant Features
Regression
Error for Regression Model
Text

Build Dataset

Web Crawling the links of the different movies from each website. Then scraping the data fields for each webpage of the movie. And also scraping the business page of each movie and only collecting information for USA only.

Merging the fields into one database..

Regression

1. We will use the collected features of the movie data.
2. We will use linear regression model to predict the gross revenue based on features like budget.number of screens,opening week revenue,critic rating,etc.
3. We will analyse the relationship of these features with gross total revenue.
4. We will make a model that will adapt to the type of movie that is genre specific model .
5. Model will select features based on genre and predict accordingly.

Selection of Features using Pearson Correlation,Recursive Feature Elimination and Regularization Models

Clean Data

- Removing Movies with no revenue data
- Bringing the data to proper format .
- Remove inconsistent data.
- Currency Conversions , converting all revenue data to US Dollars($).
- Splitting the data by year of release
- Converting revenue data to numeric fields.

Inflation

Observation was that movie ticket prices have changed over the years. We got the movie ticket price change from Wikipedia and using this change as parameter we found the inflation adjusted revenue data for each movie.

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(Source :Wikipedia)

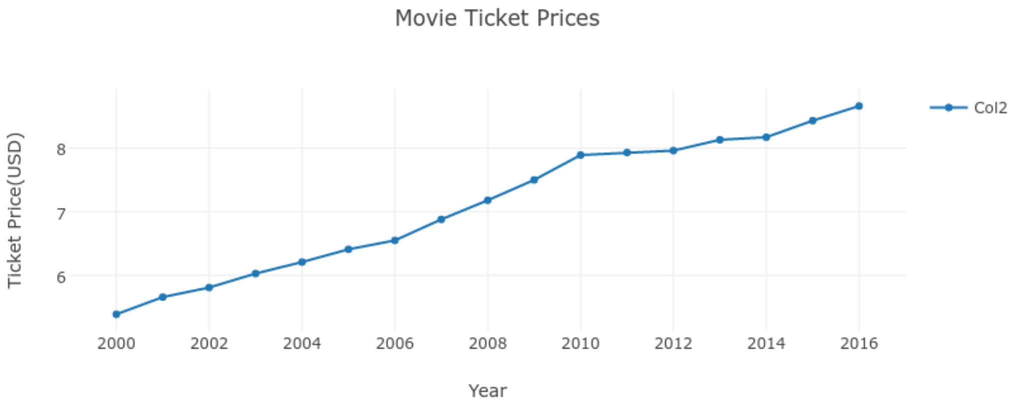


Figure 7.1: Year vs Movie Ticket Prices(Source :Wikipedia)

7.3.1 Standardization

Data collected about the movies like budget,opening weekend revenue, all user and critic ratings are all having different ranges.To use them together in a function we have to bring them to a same level or scale otherwise algorithm won't work properly.This process is called Feature Scaling.

This will make all the different features having different ranges to contribute proportionally to the final gross revenue.In fact the ranges are unbounded that budget of the movie or amount of gross revenue has no maximum value.After doing this all features will have zero mean and unit variance.

Reasons for using Standardization:

1. There is certain minimum and maximum value.
2. To handle outliers.

Method:

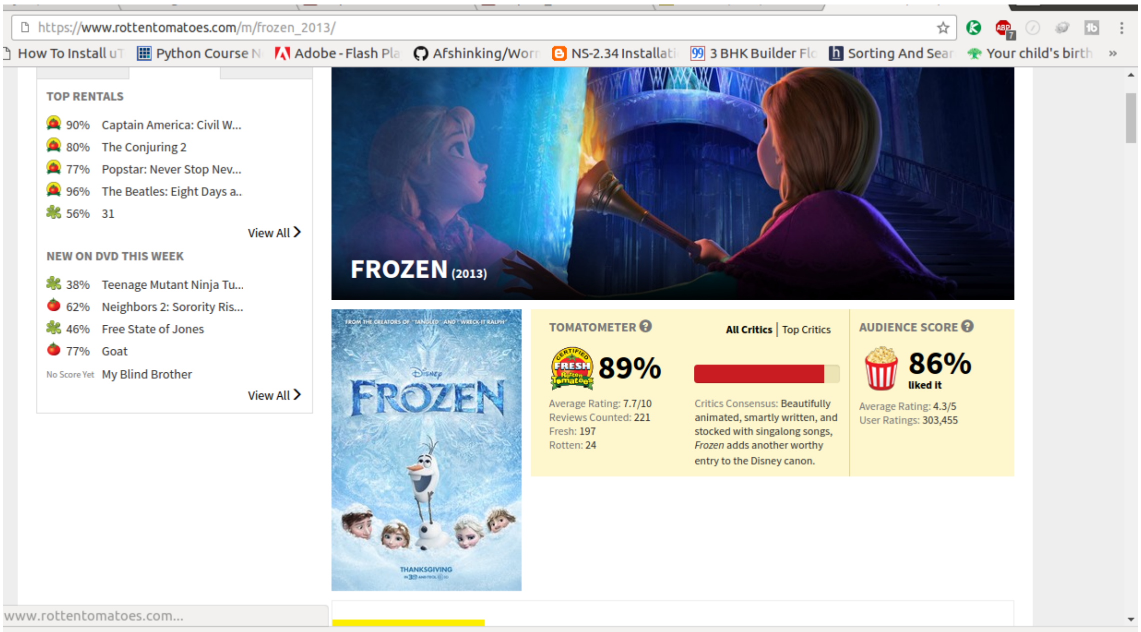
1. Calculate mean and standard deviation for each feature.
2. Subtract the mean from each feature.
3. Then divide the resultant value for each feature by its standard deviation.

Do this for each sample and feature:
Formula for Standardization is :

$$x_1 = \frac{x - mean}{Standarddeviation} \tag{7.1}$$

Feature	Mean	Standard Deviation
Budget	53834656.126 USD	54320809.576 USD
Opening Weekend	17763934.998 USD	25745328.588 USD
Screens	1835.06	1367.26
MetaScore(Out of 100)	54.05	18.29
Tomato Meter(Out of 100)	53.46	28.14
Tomato Rating(Out of 10)	5.70	1.49
User Meter(Rotten)(Out of 100)	61.44	18.53
User Rating(Rotten)(Out of 5)	3.36	0.452
User Rating(IMDB)(Out of 10)	6.41	1.07
Popularity(IMDB)	2236.0	1337.78
Popularity(Rotten)	573001.0	3605809.2

Table 7.2: Statistics (Features)



TOP RENTALS

90%	Captain America: Civil W...
80%	The Conjuring 2
77%	Popstar: Never Stop Nev...
96%	The Beatles: Eight Days a..
56%	31

View All >

NEW ON DVD THIS WEEK

38%	Teenage Mutant Ninja Tu...
62%	Neighbors 2: Sorority Ris...
46%	Free State of Jones
77%	Goat
No Score Yet	My Blind Brother

View All >

TOMATOMETER ?

All Critics | Top Critics

89%

Average Rating: 7.7/10
Reviews Counted: 221
Fresh: 197
Rotten: 24

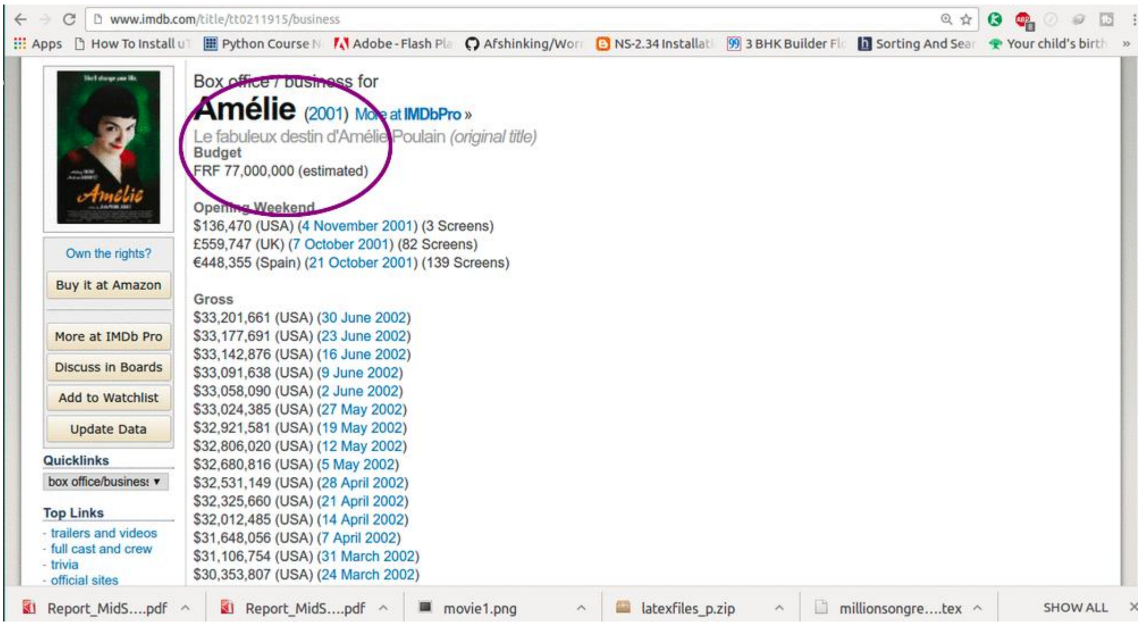
Critics Consensus: Beautifully animated, smartly written, and stocked with singalong songs, *Frozen* adds another worthy entry to the Disney canon.

AUDIENCE SCORE ?

86%
liked it

Average Rating: 4.3/5
User Ratings: 303,455

Figure 7.2: Scraping From Rotten Tomatoes Page(Source:Rotten Tomatoes)



Box office / business for

Amélie (2001) More at IMDbPro »

Le fabuleux destin d'Amélie Poulain *(original title)*

Budget
FRF 77,000,000 (estimated)

Opening Weekend
$136,470 (USA) (4 November 2001) (3 Screens)
£559,747 (UK) (7 October 2001) (82 Screens)
€448,355 (Spain) (21 October 2001) (139 Screens)

Gross
$33,201,661 (USA) (30 June 2002)
$33,177,691 (USA) (23 June 2002)
$33,142,876 (USA) (16 June 2002)
$33,091,638 (USA) (9 June 2002)
$33,058,090 (USA) (2 June 2002)
$33,024,385 (USA) (27 May 2002)
$32,921,581 (USA) (19 May 2002)
$32,806,020 (USA) (12 May 2002)
$32,680,816 (USA) (5 May 2002)
$32,531,149 (USA) (28 April 2002)
$32,325,660 (USA) (21 April 2002)
$32,012,485 (USA) (14 April 2002)
$31,648,056 (USA) (7 April 2002)
$31,106,754 (USA) (31 March 2002)
$30,353,807 (USA) (24 March 2002)

Figure 7.3: Currency Conversions and Dealing with unicode characters(Source :IMDB)

Chapter 8

Estimation Using Regression Models

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. Let A denote the actual value and F denote the predicted value. Let n number of test movies:

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{8.1}$$

8.2 Multiple Linear Regression

Assuming linear relationship between gross revenue and other features, we thought of using multiple linear regression.

Let us use all the features into our model:

$$Y = W1 * Budget + W2 * OpenWeek + W3 * Screens + W4 * userRating + .. + b \tag{8.2}$$

In [10] , multiple linear regression was used and MAPE error was around 130 percent.

Approach	MAPE
Linear Regression	110

Table 8.1: Linear Regression(All movies)

8.3 Split on the Basis Of Genre

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.

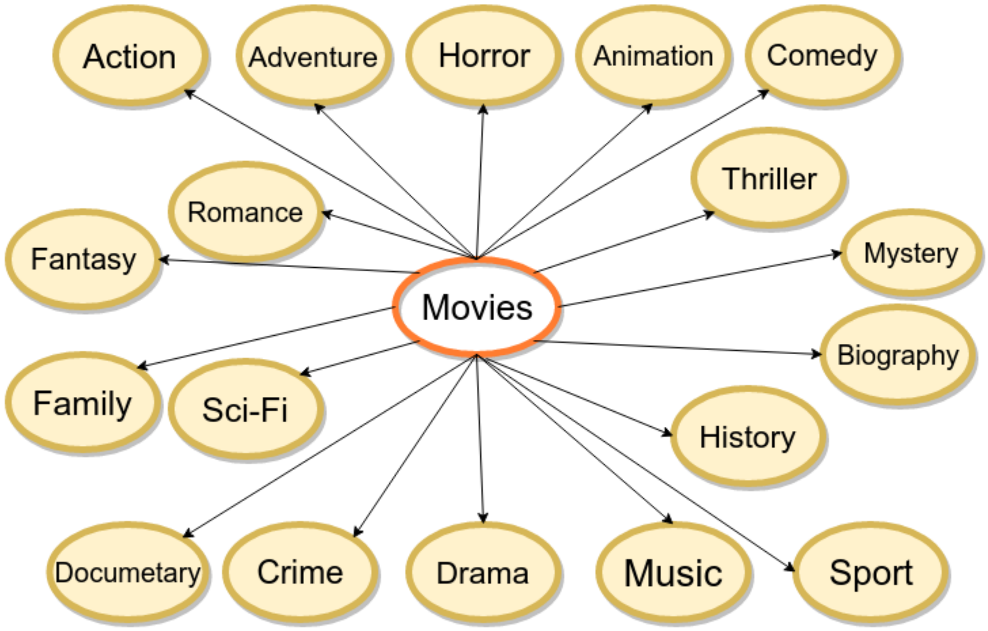


Figure 8.1: Split on the Basis Of Genre of Movie

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 on the basis on genre.
2. We then used feature selection algorithms for finding out the relevant features for each genre.
3. Thus finding out the best combination of features for each genre.
4. We then used multiple linear regression as our model.
5. Testing is done using 10 fold cross validation and finding out the mean value of error for 10 folds.
6. This way we are making a model that will change its set of input parameters required based on genre of movie .
7. Calculate error for each genre of movie for its best combination.

Chapter 9

Feature Selection

Feature Selection is used basically for two reasons:

1. To avoid over fitting by reducing number of features and to improve generalization of model.
2. To gain better understanding of features and their relationship to response variable.

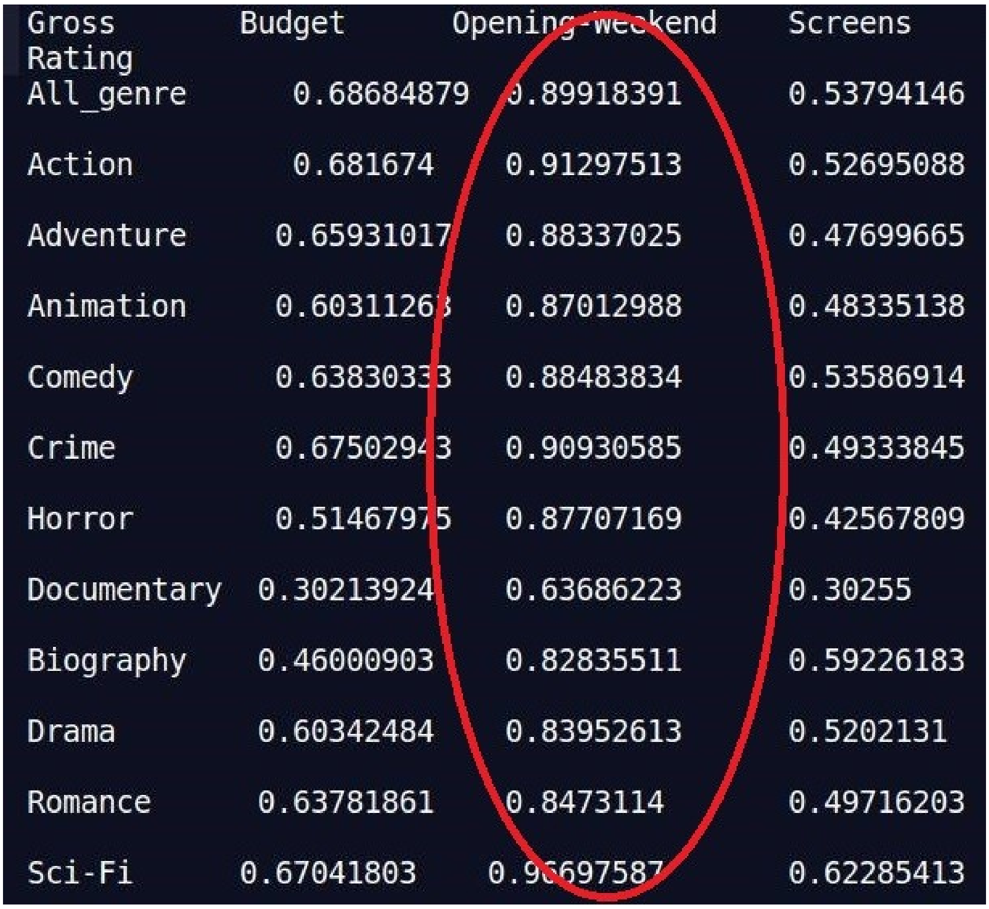
Methods we used are:

1. Uni-variate Feature selection (Pearson correlation)
2. RFE (Recursive Feature Selection)
3. Best subsets regression method

9.0.1 Pearson correlation

In this each feature is evaluated independently with respect to the gross revenue.For understanding the relationship between different features and gross revenue. We calculated the correlation coefficient for each genre of the movie. Opening Weekend has the strongest effect on the Gross revenue.We considered Opening Weekend to be included always in the feature set. Every genre showed correlation value of above 0.6.

This just give us a general idea about how features affect the value of gross revenue.Since the features also depend on each other and there can non linear relationships. We can't depend on this correlation's result.

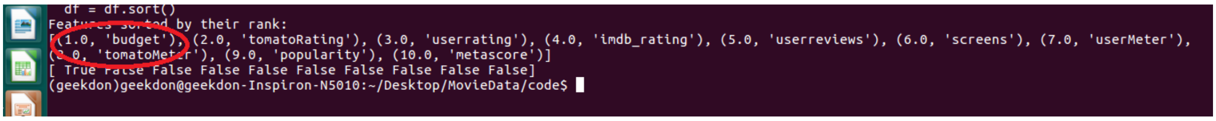


Gross Rating	Budget	Opening-weekend	Screens
All_genre	0.68684879	0.89918391	0.53794146
Action	0.681674	0.91297513	0.52695088
Adventure	0.65931017	0.88337025	0.47699665
Animation	0.60311263	0.87012988	0.48335138
Comedy	0.63830333	0.88483834	0.53586914
Crime	0.67502943	0.90930585	0.49333845
Horror	0.51467975	0.87707169	0.42567809
Documentary	0.30213924	0.63686223	0.30255
Biography	0.46000903	0.82835511	0.59226183
Drama	0.60342484	0.83952613	0.5202131
Romance	0.63781861	0.8473114	0.49716203
Sci-Fi	0.67041803	0.96697587	0.62285413

Figure 9.1: Pearson Correlation Values

9.0.2 Recursive Feature Elimination

It repeatedly constructs the regression model and dropping the worst performing feature with the least weight at each step until we are left with number of desired feature.Using this type of method we may drop features which may have good relation with the dependent variable but were suppressed by the presence of other features.



```
  df = df.sort()
Features sorted by their rank:
[(1.0, 'budget'), (2.0, 'tomatoRating'), (3.0, 'userrating'), (4.0, 'imdb_rating'), (5.0, 'userreviews'), (6.0, 'screens'), (7.0, 'userMeter'),
(8.0, 'tomatoMeter'), (9.0, 'popularity'), (10.0, 'metascore')]
[ True False False False False False False False False]
(geekdon)geekdon@geekdon-Inspiron-N5010:~/Desktop/MovieData/code$ 
```

Figure 9.2: Recusrive Feature Elimination For Comedy Movies

9.0.3 Best subsets regression

Therefore , we used best subset regression which in most of the cases where the number of features are less performs best.Here we have 11 features in the dataset.Best subsets regression is an procedure which identifies the best subset or best fitting set of features based on some statistical criteria.There can be different statistical criteria like Mean Square Error, R squared,etc. Here, we will be using Mean Absolute Percentage Error(MAPE).

Hence, we will say the best combination or subset of features is the one with the least MAPE.

9.0.4 Best Combination for Each Genre

Genre	Best Combination
Action	Opening Weekend,Popularity(IMDB)
Adventure	Opening Weekend,Budget
Animation	Opening Weekend,Budget
Drama	Opening Weekend,Budget,Popularity(IMDB)
Comedy	Opening Weekend,Budget
Sci-Fi	Opening Weekend
Romance	Opening Weekend,Budget,Popularity(Rotten)
Music	Opening Weekend,Budget,UserMeter,Popularity(Rotten)
Fantasy	Opening Weekend,Budget,Screens,Popularity(IMDB),UserRating
History	Opening Weekend,Popularity(IMDB),TomatoRating
Documentary	OpenWeekend,Popularity(IMDB),TomatoRating,UserRating,Popularity(Rotten)
Horror	Opening Weekend,Popularity(Rotten)
Mystery	Opening Weekend,Budget,Popularity(IMDB),Popularity(Rotten)

Table 9.1: Best Combination for Each Genre

From the above table we can say that:

1. Shows that 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 and Popularity don't seem to affect the performance of a lot of movies. Only Documentary,Drama,History,Music were affected a bit.

4. Budget and Opening Weekend seem to make the best combination.

9.1 Genre wise Analysis

We ran the Multiple Linear Regression Model for each genre of the movie. For testing we used 10 fold Cross validation and took the mean of the MAPE(Mean absolute Percentage Error) for error of each fold.

Genre	MAPE
Action	46.45
Adventure	49.16
Animation	36.693
Drama	95.45
Comedy	50.26
Sci-Fi	24.20
Romance	90.45
Music	60.70
Fantasy	47.90
History	62.38
Documentary	42.57
Horror	23.47
Mystery	49.89

Table 9.2: MAPE for each genre

9.2 Multi Genre Analysis

We found that result improved when used the linear regression model genre wise.Now to deal with movies containing more than one genre.

We will do the following things:

1. First we will run the linear model for each genre.
2. Testing is done using 10 fold cross validation.
3. Model will use the best combination for that genre to predict the gross revenue for that genre.
4. Once we find the gross revenue values corresponding to each genre , we will take the mean of these.
5. That would be our estimated Gross Revenue.

Approach	MAPE
Linear Regression	53.966

Table 9.3: Linear Regression(All movies)

9.3 Split of Training Set on the Basis of Critic Rating and Screens

After splitting the dataset on the basis of genre,we split it on the basis of Critic rating and Number of Screens.

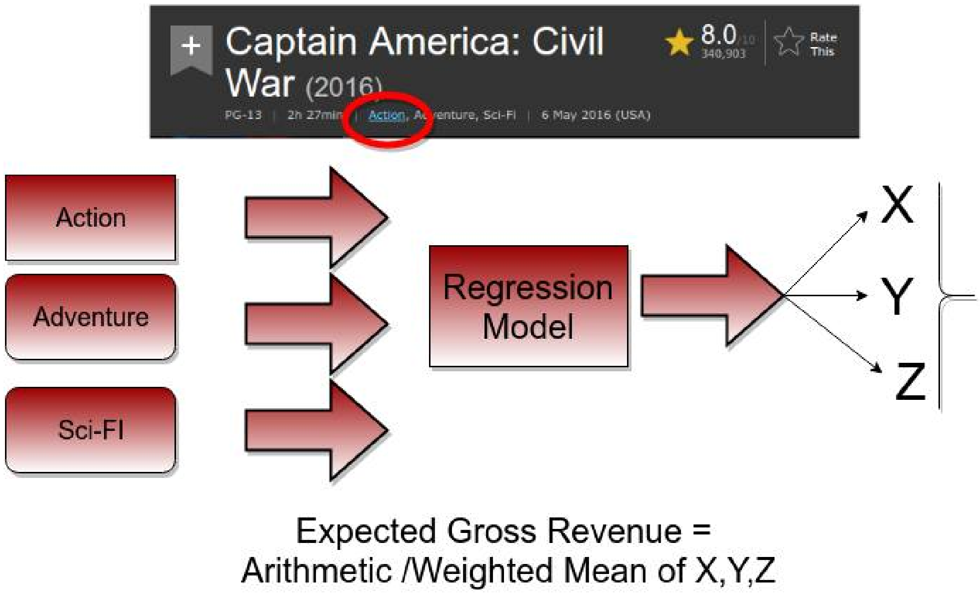

+
Captain America: Civil War (2016)
8.0/10
340,903
Rate This
PG-13 | 2h 27min | Action, Adventure, Sci-Fi | 6 May 2016 (USA)
Action
Adventure
Sci-FI
Regression Model
X
Y
Z
Expected Gross Revenue =
Arithmetic /Weighted Mean of X,Y,Z

Figure 9.3: Multi Genre Method

For Critic rating, we observed there are two type of movies(Type1 and Type2).
Type1, High Critic Rating > 5.

Type2, Low Critic Rating <= 5.

For Number Of Screens, we observed there are two type of movies(Type3 and Type4).

Mean value of Screens Taken from the Training Set is 1835.
Type3, High Critic Rating > 1835.

Type4, Low Critic Rating <= 1835.

From now:
Training Data: Movies from the Year 2000 to 2013
Test Data: Movies from the year 2014 to 2015 (400 in number)
Now when a test movie comes, if it lies in type1 then it will use that training set to predict the revenue.Similarly, for all types.

Split	All	High	Low
Rating	34.18(400)	42.6(244)	24.25(156)
Screens	39.92(400)	26.96(210)	127.32(190)

Table 9.4: Linear Regression Result

We compared with [10] , they got 262 percent MAPE in Low Screen Movies and 37.6 in High Number of Screens.

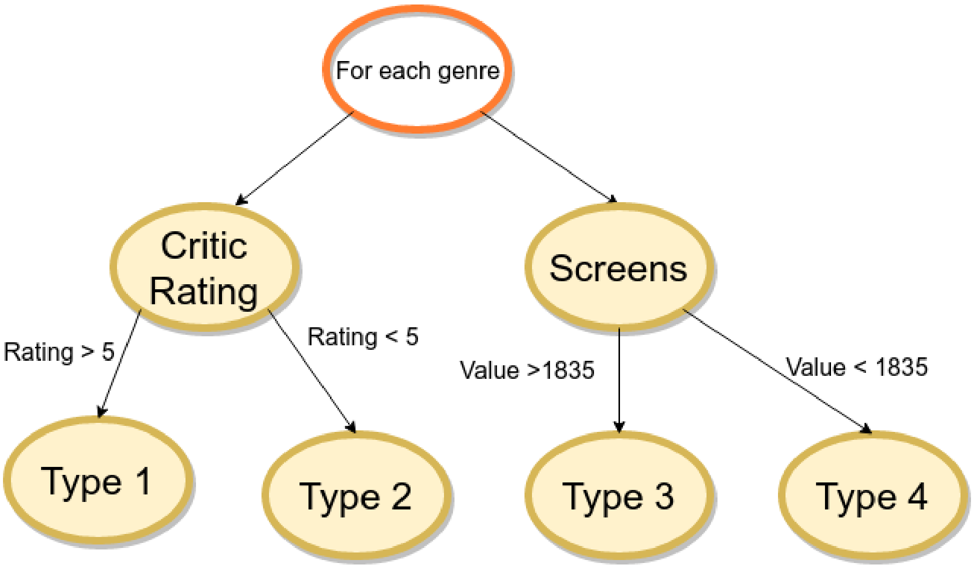


Figure 9.4: Splitting on the Basis of Critic Rating(TomatoRating) and Screens

9.4 Local Regression Models

9.4.1 Neighbour Search

For each test movie, we will try to find the nearest data items or training items. We made two vectors based on the type of genre, Like for Adventure movies, (Opening Weekend,Budget) will be the vector. Now, we calculated the euclidean distance between the test vector and each of the training data vector, we sort the distances and picked up the 50 nearest ones. We now used only these for the training of our prediction model.

9.4.2 Linear Regression

Now with hope of further improvements we used local regression models. We now train our multiple linear Regression model on these 50 neighbors and predict for the test data of 400 movies.

9.4.3 Decision Tree Regression

It breaks down the dataset into smaller and smaller subsets with decision nodes and leaf nodes.Leaf nodes contains small subset of dataset. We find the mean value of gross revenue for each leaf node using the samples in the leaf. We ran the decision tree regression algorithm and trained using the 50 neighbors.

Algorithm	All(400)	High(210)	Low(190)
Linear Regression	37.966	18.6	84.9
Decision Tree Regression	25.77	11.8	50.6

Table 9.5: Local Models(On the Basis of Splitting by Screens)

Algorithm	All(400)	High(244)	Low(156)
Linear Regression	32.47	41.38	23.21
Decision Tree Regression	28.78	29.17	17.04

Table 9.6: Local Models(On the Basis of Splitting by Critic Rating(TomatoRating)

By Observing the above tables,we see that model performs poorly on movies with low number of Screens (that is less than 1835).To overcome we thought of combining the both splits, we see that error when split is done on the basis of rating is always less compared to 50 percent error on low number of Screen movies.

When the movie is released on screens > 1835, model should use the Type3 training set.

When the movie is released on screens < 1835, model should use Rating Split and use Type1 or Type2 training Set accordingly.

The 156 test items with low number of screens were tested with Type1 and Type2 and error reduced to 28.57 percent.

Algorithm	MAPE
Decision Tree Regression	24.76

Table 9.7: Local Models(On the Basis of Splitting by Rating and Screens)

Chapter 10

Result

We tested the 400 movies of the year 2014 to 2015.We found out that after splitting the dataset on the basis of genre and doing further analysis on the basis of Number Of Screens. If number of screens are less than Mean value 1835, we should use the split on the basis of critic Rating and then predict the gross revenue.Error for Screens released in more than 1835 screens(210 test movies) is as low as 11.8 percent. It seems logical too ,as the movies which are released in large number of theatres earn good amount of income whereas those which are released at less places would earn good if and only if they are good or critics have rated them good.Using this combined type of approach ,Error(MAPE) for 400 test movies is 24.76 percent.

Algorithm	MAPE
Decision Tree Regression	24.76

Table 10.1: Local Models(On the Basis of Splitting Of Rating and Screens)

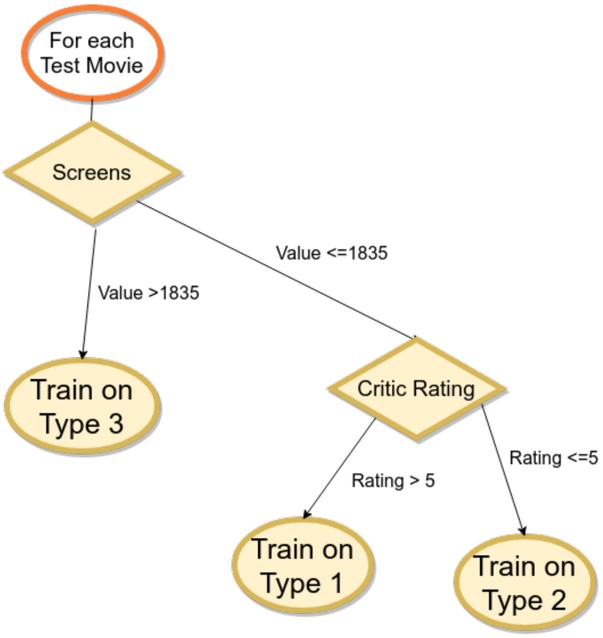


Figure 10.1: Final Approach

Chapter 11

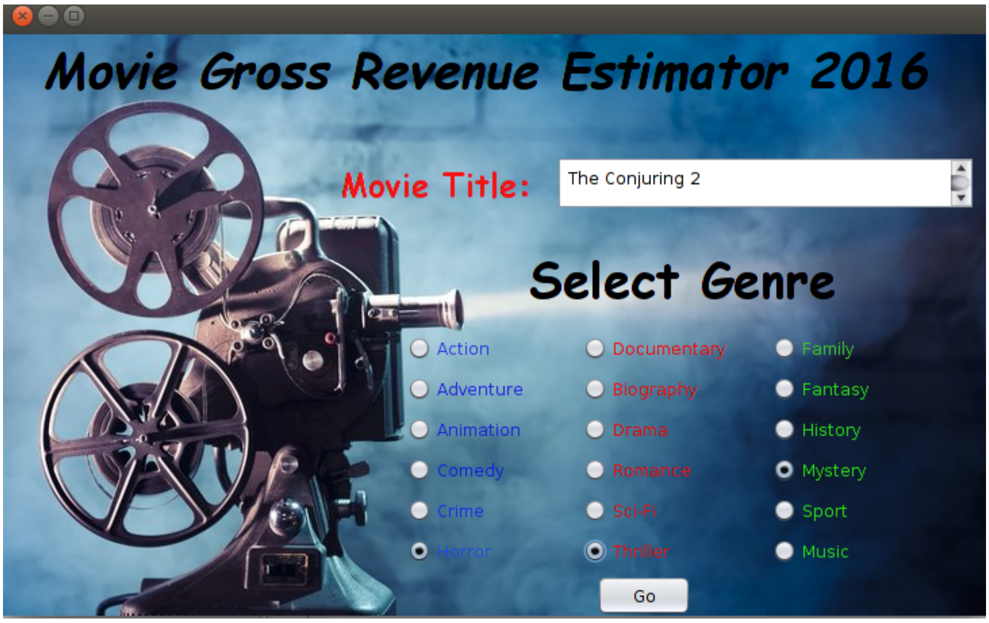
GUI

We created a GUI application for estimating the gross revenue for a movie.

Working:

1. First select the genres of the movie
2. Fill in the required features for the movie
3. Get the Estimated Revenue

Snapshots:



Movie Gross Revenue Estimator 2016

Movie Title: The Conjuring 2

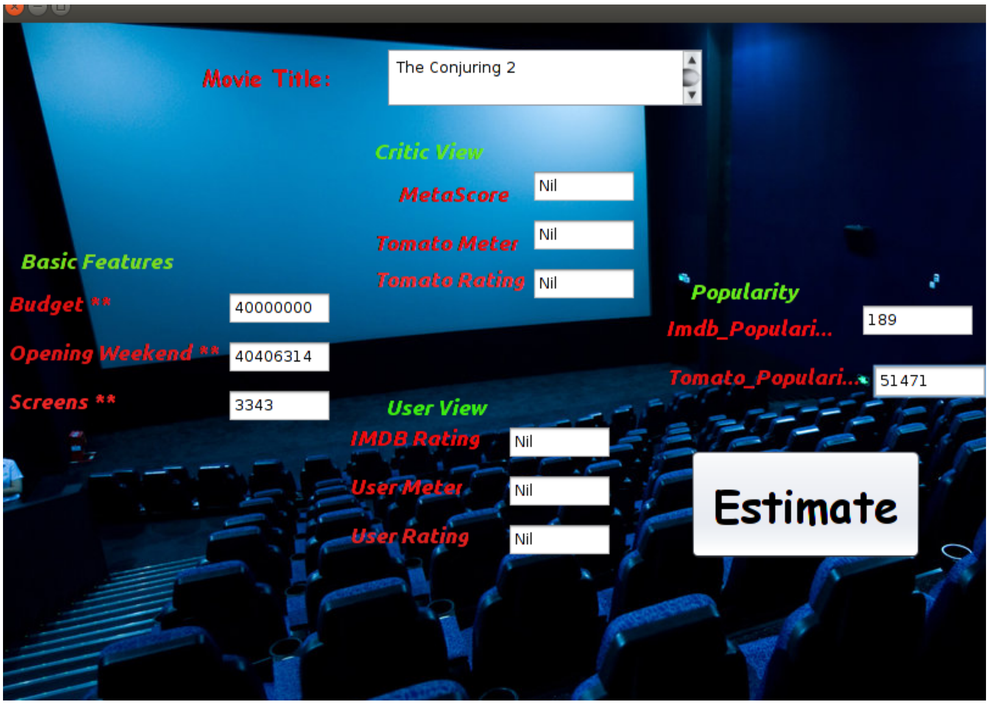
Select Genre

○ Action
○ Adventure
○ Animation
○ Comedy
○ Crime
● Horror

○ Documentary
○ Biography
○ Drama
○ Romance
○ Sci-Fi
● Thriller

○ Family
○ Fantasy
○ History
● Mystery
○ Sport
○ Music

Figure 11.1: Java Application for Estimating Revenue



Movie Title: The Conjuring 2

Critic View

MetaScore: Nil

Tomato Meter: Nil

Tomato Rating: Nil

Basic Features

Budget **: 40000000

Opening Weekend **: 40406314

Screens **: 3343

Popularity

Imdb_Populari...: 189

Tomato_Populari...: 51471

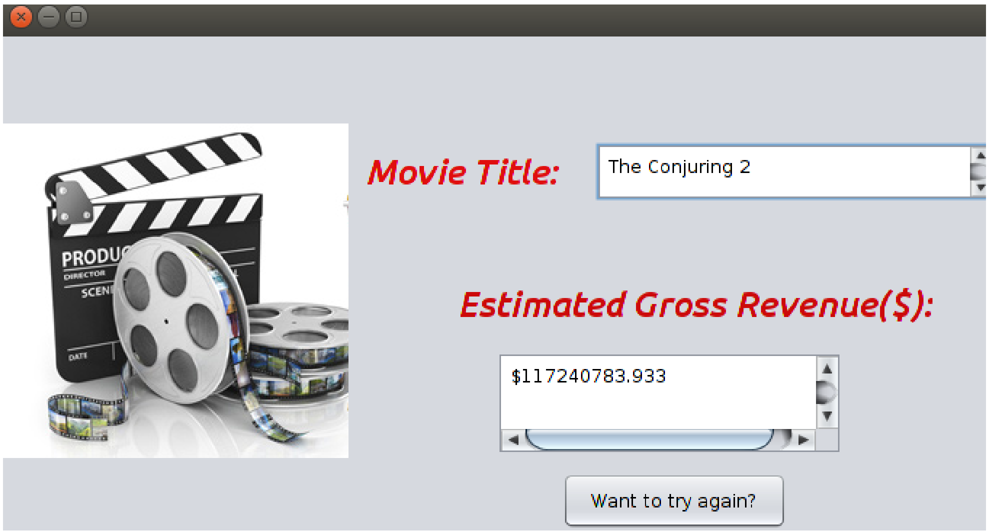
User View

IMDB Rating: Nil

User Meter: Nil

User Rating: Nil

Figure 11.2: Fill in the required features



Movie Title: The Conjuring 2

Estimated Gross Revenue($): $117240783.933

Figure 11.3: Estimated Revenue

Chapter 12

Test On Recent Releases of 2016

Name	Real Gross	Predicted Gross	MAPE(Percent)
Conjuring 2	102,461,593 USD	117,240,783 USD	14.42
The Angry Birds	107,506,776 USD	145,956,013 USD	35.5
Deadpool	363,024,263 USD	358,836,741 USD	1.15
The Legend Of Tarzan	126,585,313 USD	113,305,012 USD	10.49
The Jungle Book	363,995,937 USD	425,317,712 USD	16.8

Table 12.1: Test on Recent 5 Releases of 2016

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

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

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

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

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

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

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