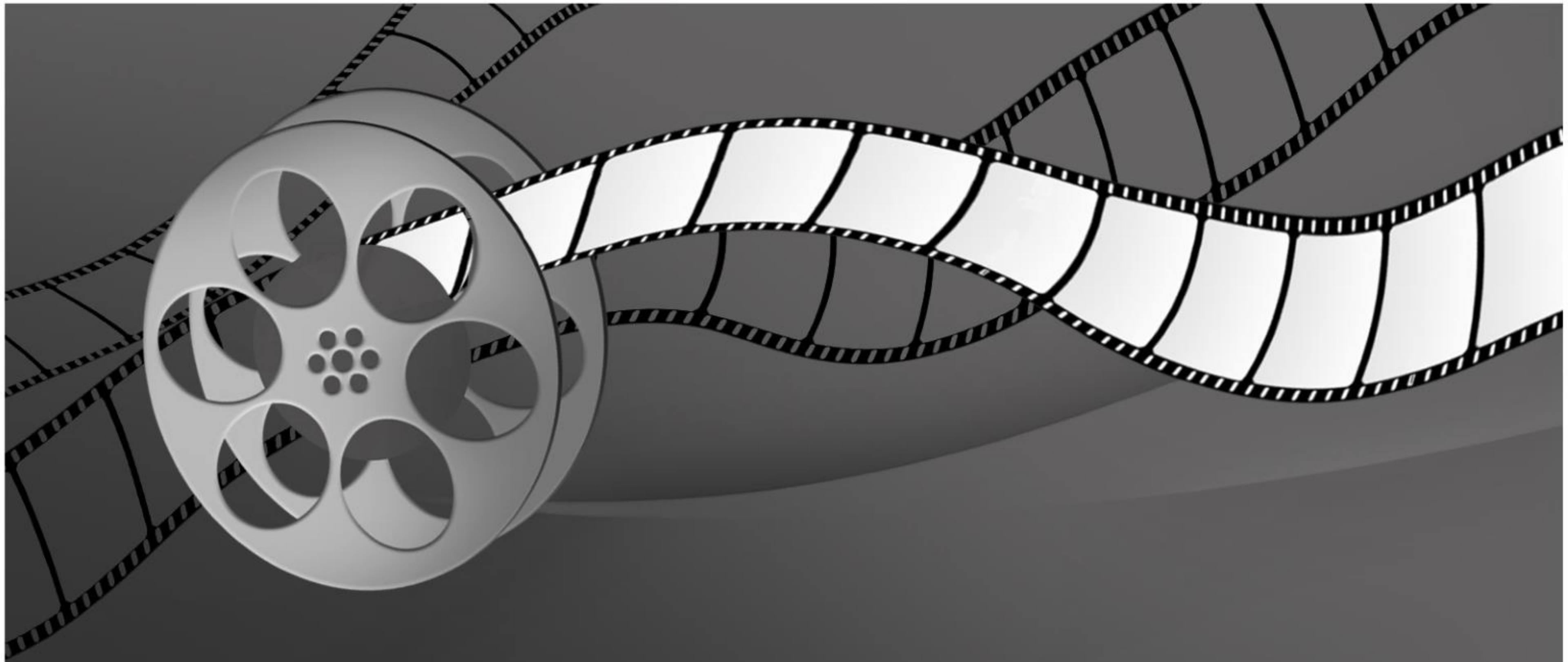


Movie Gross Revenue Estimation

VII Semester Project

Project Mentor and Guide:

Dr. Sonali Agarwal

Team Members

- ✓ Abhishek Kumar Agrawal(IIT2013128)
- ✓ Shubham Bhendarkar (IIT2013172)
- ✓ Shaiwal Sachdev (IIT2013196)

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.

Motivation

This model is useful for :

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

Data Set

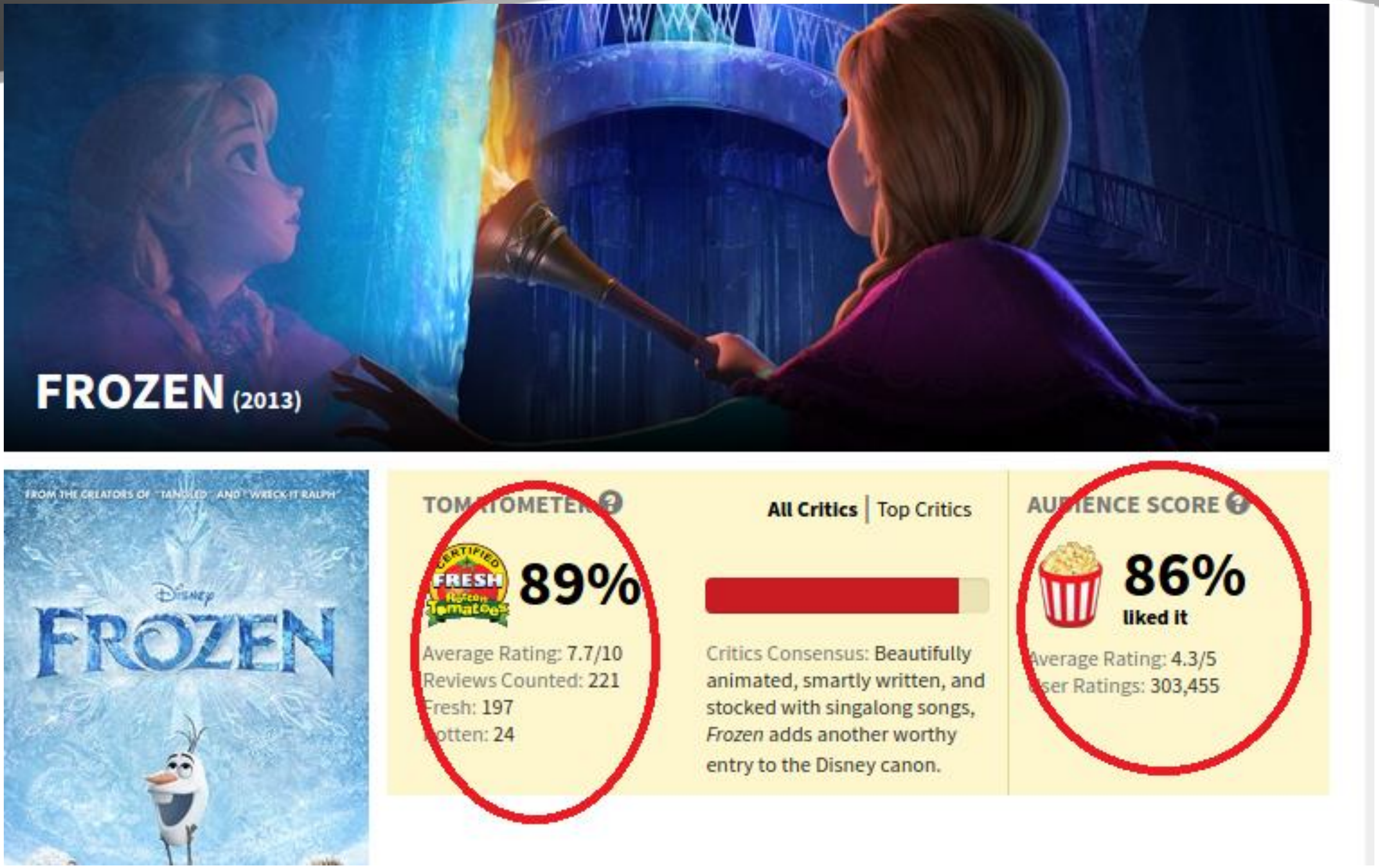
- ✓ Data is collected through crawling and scraping movie webpages of websites like IMDB, rottentomatoes, Metacritic etc.
- ✓ Revenue data is collected from IMDB Business page of each movie.
- ✓ All the revenue data is of USA region only.
- ✓ We collected the data for movies released from year 2000 to 2015 only.

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

IMDB

Frozen (2013)
7.6/10
425,645
Rate This
PG | 1h 42min | Animation, Adventure, Comedy | 27 November 2013 (USA)
FROZEN
2:30 | Trailer
7 VIDEOS | 136 IMAGES
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)
nations. See more awards »

ROTTEN TOMATOES



IMDB Business Page

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

Own the rights?

Buy it at Amazon

More at IMDb Pro

Discuss in Boards

Add to Watchlist

Update Data

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)

$400,7[illegible] (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)

Fields Information

Movie Features:
Title of the movie
Genre of the movie
Release Date
Total Runtime (in minutes)
Year of Release
MPAA Rating

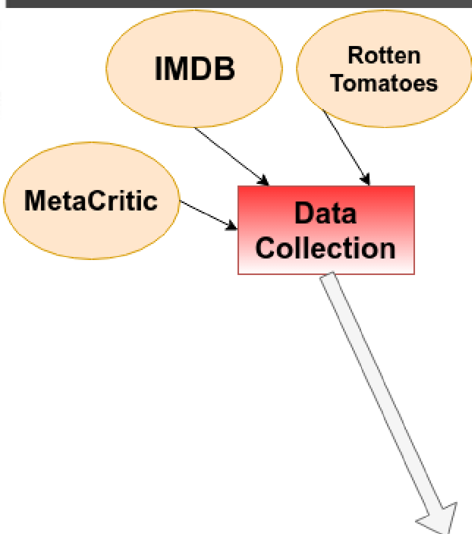
Important Features
Budget Of the Movie (USD)
Opening Weekend Revenue (USD)
Number Of Screens in Opening Weekend
Movie Gross Domestic Revenue (USD)

Critic View
From Rotten Tomatoes(Tomatometer, Tomato Rating)
From Metacritic (Metascore)

User View
From Rotten Tomatoes (User Meter, User Rating)
From IMDB (IMDB User Rating)

Popularity
From IMDB
From Rotten Tomatoes

Proposed Approach



Build Dataset

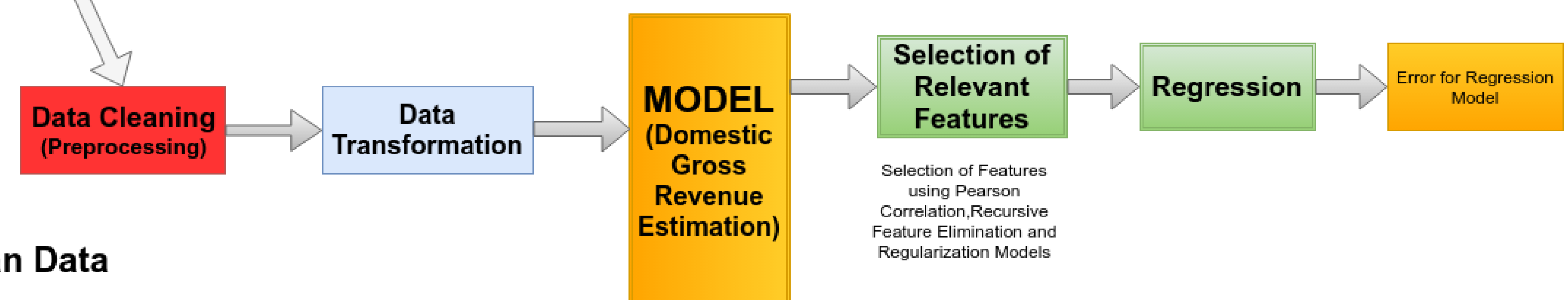
Web Crawling the links of the different movies from each website. Then scraping the data fields for each webpage of the movie. And also scraping the business page of each movie and only collecting information for USA only.

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.

Process Flow



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.

Text

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.

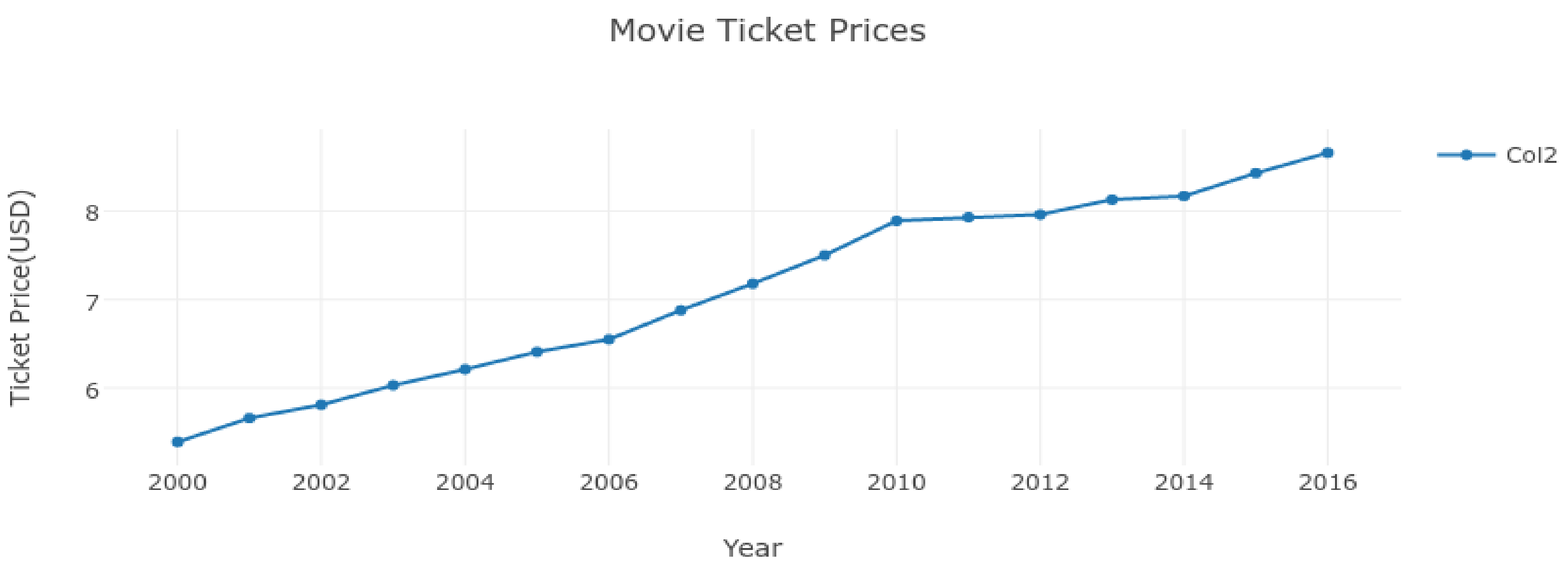
4. Removing inconsistent data.

5. Budget of the movie was in different currencies . Currency Conversions had to be done for converting all this data to US Dollars.

6. Converting revenue data to numeric fields.

Data Transformation

- Due to inflation price of tickets have changed so much over the years. We used this change in ticket price and as inflation parameter and calculated the inflation adjusted revenue data for each movie. Current Ticket Price in 2016 is 8.66 USD.



Source : Wikipedia

Standardization

- This method is used to standardize the range of independent variables or features of data.
- Minimum and maximum values of the features (Budget, Screens, Opening Weekend , gross revenue etc) are unbounded.
- Therefore for handling such types of features we used Standardization.
- It will make the values of each feature in the data to have zero mean and unit variance. Thus normally distributed.

Formula is :

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

Statistics

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

Estimation using Regression Models

- **Error Measurement**

Mean Absolute Percentage Error (MAPE) : It is the mean of percentage error of each sample. Let A denote actual value, Let F denote predicted value, n be the number of test movies.

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

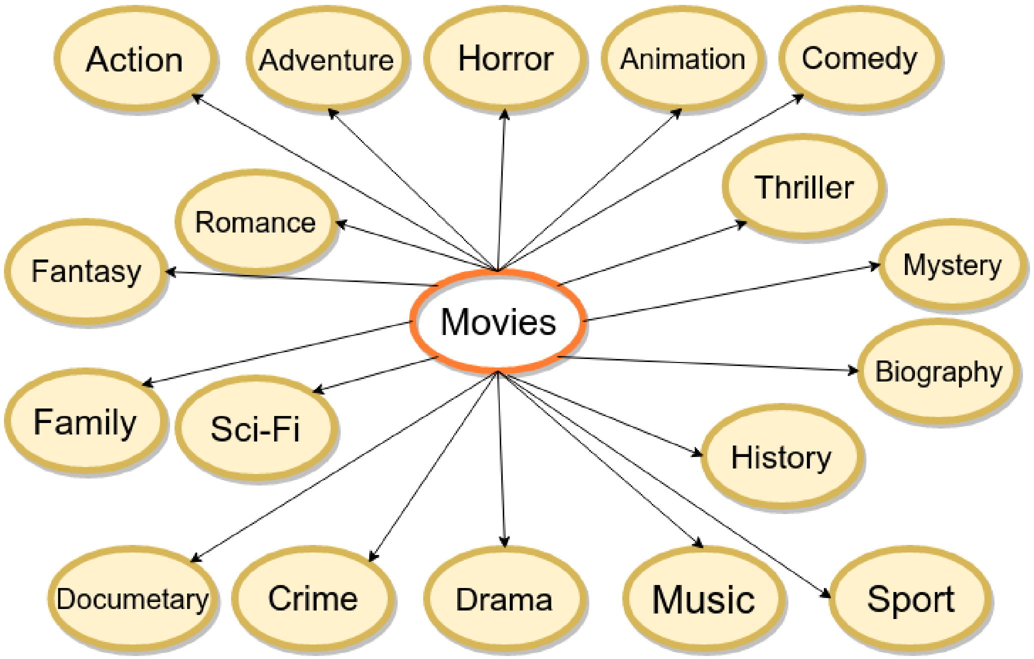
Multiple Linear Regression- We used all the features in our model.

Y = W1 * Budget + W2 * Opening Weekend + W3*Screens +.. + b.

- We calculated the error using this model, it was around **110 percent**.

Split on the basis of Genre

- ✓ As per the journal paper, different types of movies respond to different parameters differently. So, we decided to split the dataset by genre of the movie and do further analysis.



Prag, J.J. Casavant, J. 1994. An empirical study of the determinants of revenues and marketing expenditure in the motion picture industry's, Journal of Cultural Economics,

Feature Selection

- We then used feature selection algorithm, finding out the relevant features for each Genre.

Feature Selection is used mainly of two reasons :

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

Methods :

- Univariate Feature selection (Pearson correlation)
- RFE (Recursive Feature Selection)
- Best Subset Regression Method

Source : Wikipedia

- **Pearson correlation** For understanding the relationship between different features and gross revenue. We calculated the correlation coefficient for each genre of the movie.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Gross Rating	Budget	Opening-weekend	Screens	MetaScore
All_genre	0.68684879	0.89918391	0.53794146	0.12052243
Action	0.681674	0.91297513	0.52695088	0.44168585
Adventure	0.65931017	0.88337025	0.47699665	0.3711183
Animation	0.60311263	0.87012988	0.48335138	0.30223538
Comedy	0.63830333	0.88483834	0.53586914	0.10797257
Crime	0.67502943	0.90930585	0.49333845	0.10866238
Horror	0.51467975	0.87707169	0.42567809	0.18489136
Documentary	0.30213924	0.63686223	0.30255	-0.17440908
Biography	0.46000903	0.82835511	0.59226183	0.09493294
Drama	0.60342484	0.83952613	0.5202131	0.06373391
Romance	0.63781861	0.8473114	0.49716203	0.0059639
Sci-Fi	0.67041803	0.96697587	0.62285413	0.37165667

- # 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(gross revenue) but were suppressed by the presence of other features.

For Example : For Comedy Movies.

```
  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 False]
(geekdon)geekdon@geekdon-Inspiron-N5010:~/Desktop/MovieData/code$
```

Best Subset Regression

- In this we identify best subset or best fitting set of features for each genre based on some statistical criteria. Here we have used MAPE for selecting the best subset.
- This method works best when we have less number of features, (11 features)

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)

Genre Wise Result (Linear Regression)

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

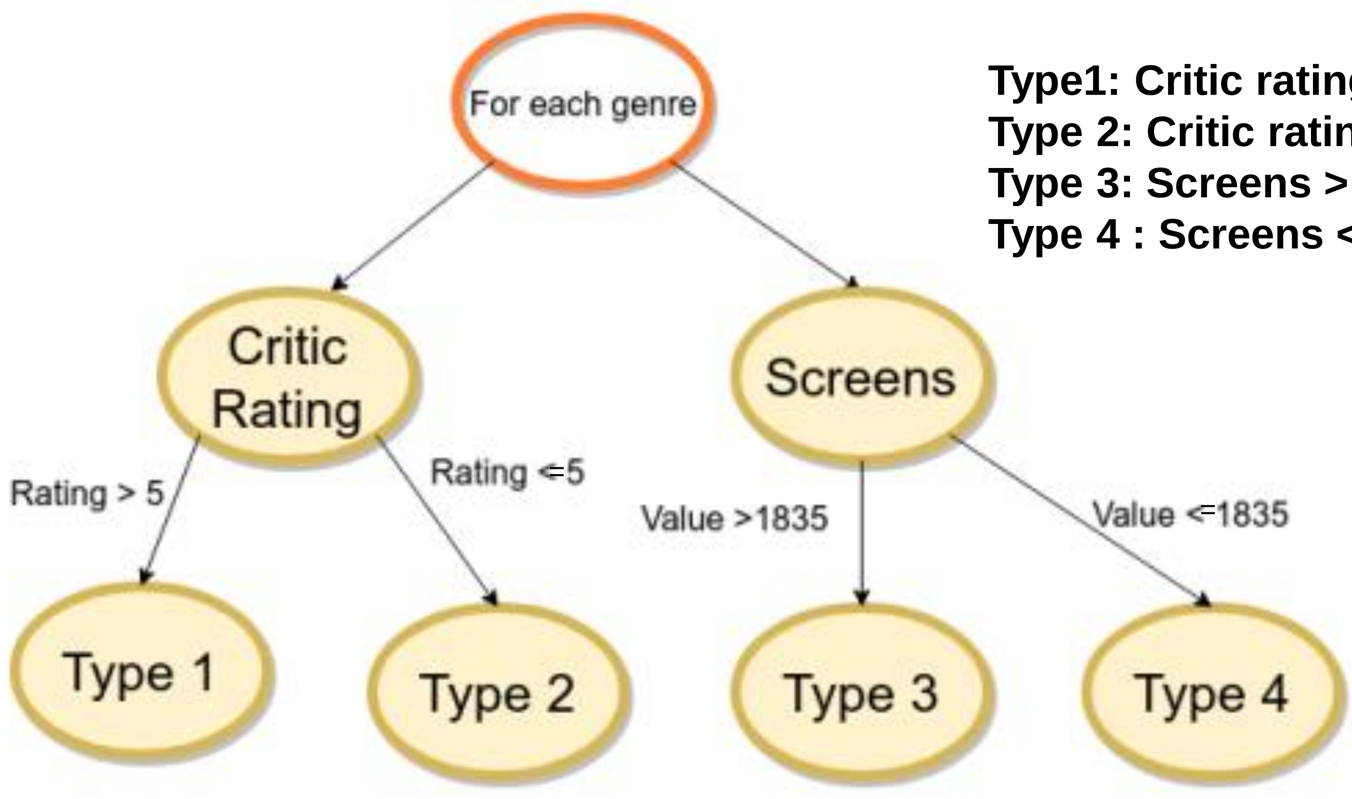
Dealing with Multi-Genre Movies

- We applied multiple linear regression for each genre using the best combination and taking the arithmetic mean of predicted values of each genre.

Approach	MAPE
Multiple Linear Regression	53.966

Split on the basis of Critic Rating (Tomato Rating) and Screens

Mean Value of Number of Screens : 1835



Type1: Critic rating > 5
Type 2: Critic rating <= 5
Type 3: Screens > 1835
Type 4 : Screens <= 1835

Split on the basis of Critic Rating and Screens

- Now test data for testing purpose we use **400 movies of the year 2014-15**
- Now when a test movie comes, we can perform testing either rating wise or screen wise.

✓ For **Rating** basis testing : if test movie has **high rating (> 5**) it will be predicted as per model trained in type 1. while if **rating is <= 5 (low)** , it will be predicted as per model trained in type 2.

✓ For **Screen** basis testing : If test movie has **number of screens > 1835**, it will be predicted as per model trained in type 3 while if it has **screens <= 1835**, it will be predicted as per model trained in type 4.

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

Multiple Linear Regression model

Local Regression Model

- **Neighbour Search** : For each test movie, we will try to find out nearest data or training items.
- Using the **best combination** as per the genre as the **feature vector**, we calculated the **Euclidean distance** between the **test** feature vector and other **training** feature vectors.
- By sorting the distances, we picked up the **50 nearest ones.**

- **Two methods:**

1 . Linear Regression =

2. Decision Tree Regression =

- We used the above two algorithms and trained them using the 50 neighbours found using neighbour search.

Local Regression Models

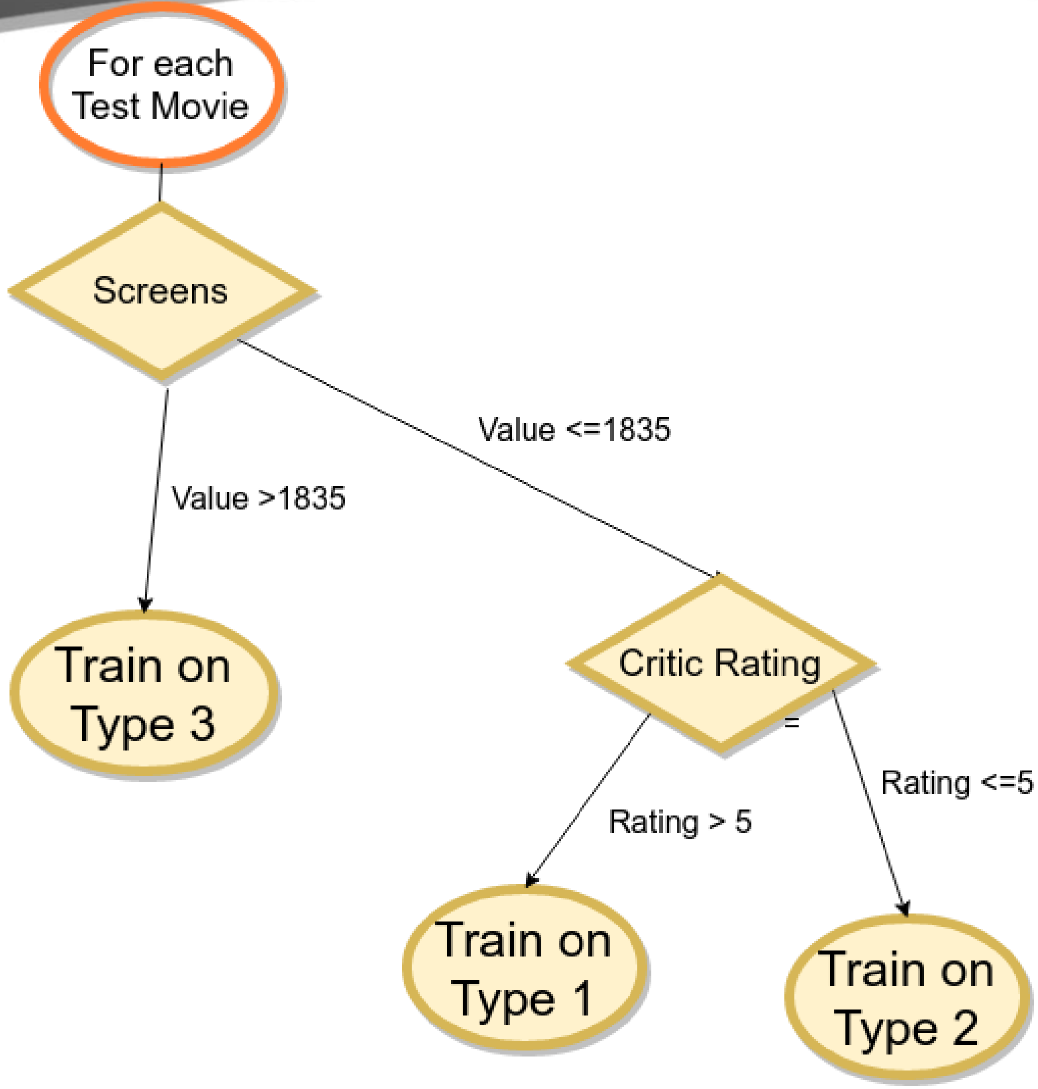
Split by Number of Screens

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

Split by Critic Rating (Tomato Rating)

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

Combined Approach



Result

- For the test movie set (190) having **lower** number of **screens** (<=1835), using the combined approach , error dropped from 50.6 % to **28.57** %.

- Now using the local decision tree regression and combined approach we tested on our test set (400).

Algorithm (local)	MAPE
Decision Tree Regression	24.76

GUI Application

Test on Recent Releases of 2016

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

THANK YOU!