Springboard Data Science Career Track Capstone 2 – Milestone report

Movie Box-office Success Prediction

Rajesh Dharmarajan 2/12/18 Springboard Career Track

TABLE OF FIGURES

1. OVERVIEW

My second Capstone project will be to predict commercial success of a movie. Movies have been one of the entertainment industries that sees huge sums of money being invested. Though making movies are art than science, this project tries to explore if there are any formulas or secret recipe in making commercially successful movies. This project will also try to explore the important parameters or factors that influence the success of a movie.

This document provides a summary of the problem, the approach taken to arrive at the solution.

1.1. PROBLEM STATEMENT

Making movies is generally considered to be more art than science. Given this perception, can the success of a movie be predicted even before it is released? Or, is movie making pure art and there will not be any correlation between the key factors like the actor, budget, director, etc. to the success of a movie.

There are huge sums of money invested in movies. According to https://www.statista.com the U.S movie box office revenue was $ 10.4 Billion in 2015. With so much money at stake, it would help to understand which variables or factors are the most important in determining the box office success.

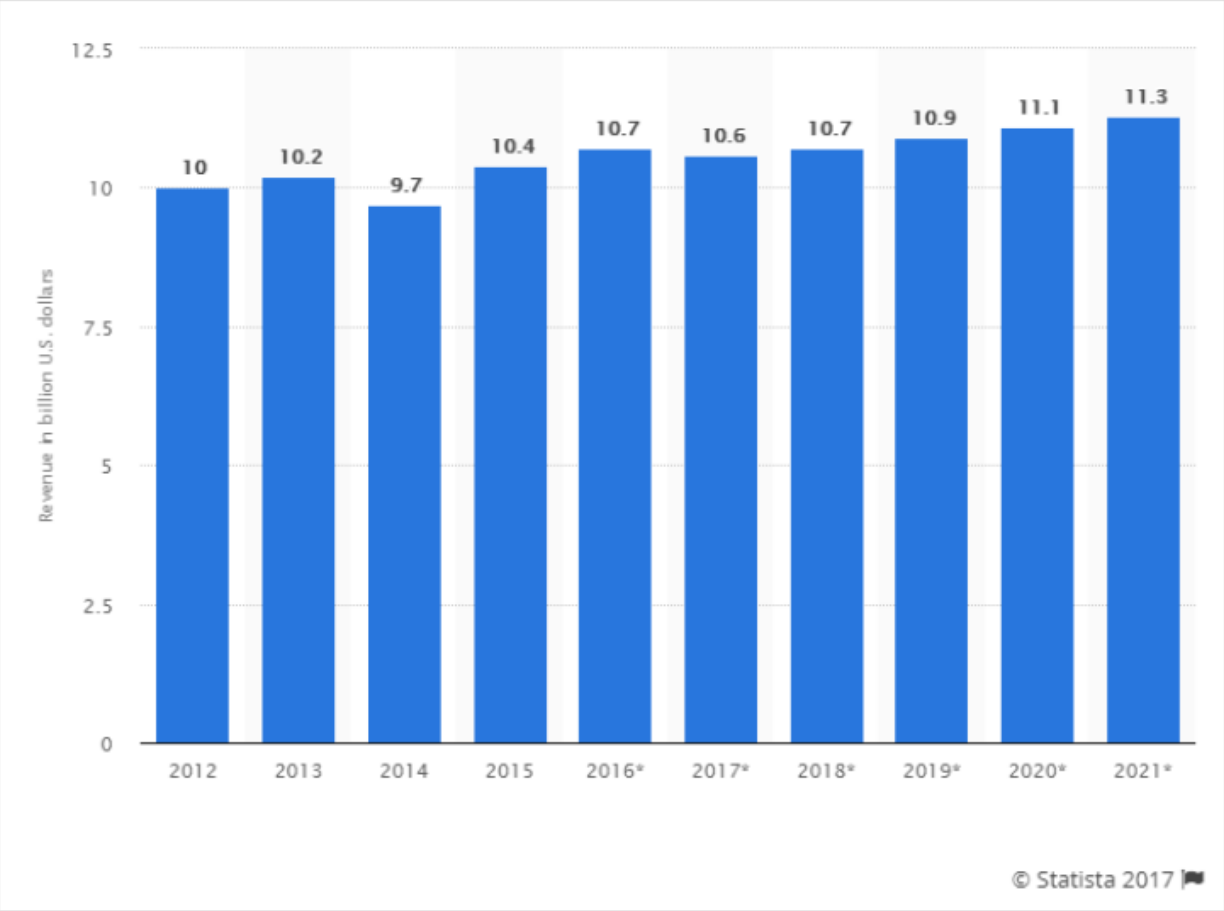

FIGURE 1 - MOVIE BOX OFFICE REVENUES

Using data about a movie that is available before the movie is released or even made, can machine learning models predict the success of a movie.

1.2. OBJECTIVE AND EXPECTED RESULTS

This project attempts to use data from existing sources like imdb.com and boxofficemojo.com to gather characteristics of a movies as it is being made or being conceived. These features could be things like the actors in the movie, the director of the movie, the season when the movie is being released, etc. This model will not use features that are created or generated after a movie is released like user reviews since the objective of this model is to forecast the success of a movie before it is released.

This model can be created as a prediction model where it will predict the box office (BO) collection of a movie or it can be created as a classification model where the degree of success can be predicted. In this project I create this model to predict the following BO collection metric as an indicator of the success / failure of the movie

$$\frac{4th\ Weekend\ GTD\ Box\ Office\ Collection - Movie\ Production\ Budget}{Movie\ Production\ Budget}$$

This model can be expanded to include more sophisticated parameters like quality of script, pre-release social media mentions, etc.

1.3. POTENTIAL CLIENTS

Potential customers for this would be

i. Movie Production Companies
ii. Actors

Anyone with a stake in the movie business or an academic interest in movies and it's making will benefit from such a model.

1.4. APPROACH

The high-level steps to solve this problem will be

1. Data cleansing and wrangling
2. Deciding metrics to be used to calculate the accuracy
3. Applying various Machine Learning (ML) algorithms to the cleansed data to find the optimal solution

The overall solution would be an iterative process, with multiple cycles of data cleansing and application of algorithms happening throughout the process until an optimal solution with satisfactory accuracy is arrived at.

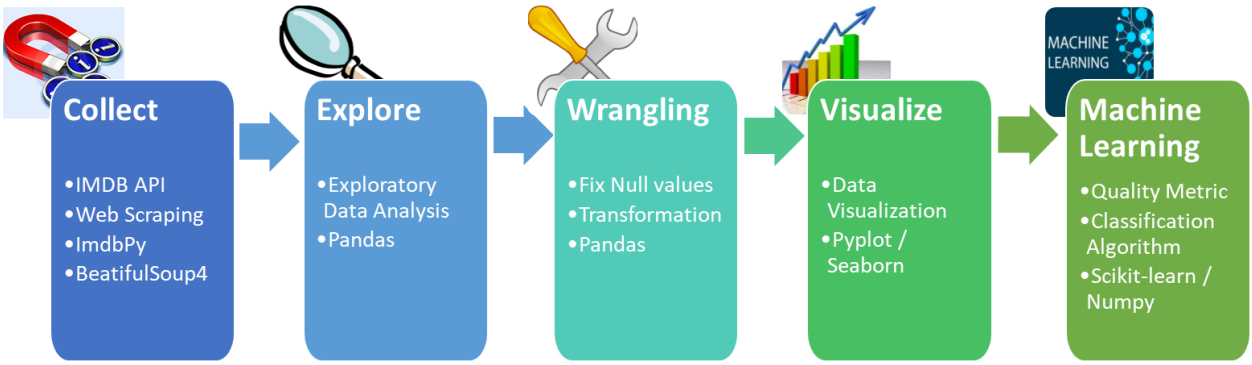

FIGURE 2 SOLUTION APPROACH

1.5. DELIVERABLES

Deliverables of this project will include

1. Code – iPython notebook. Code with relevant documentation included.
2. Report – a MS Word document that describes in detail the approach, the findings, the final result and recommendations
3. Presentation – a MS PowerPoint slides that presents the problem and results

2. About the Data

2.1. Data Used

The sources of the data to be used for this has been extracted from www.imdb.com, www.boxofficemojo.com and https://grouplens.org

Grouplens.org contains a movielens dataset that has a list of movies and their IMDB id. The list of movies has been used as a starting point to get additional data from IMDB and boxofficemojo websites. The movielens dataset from Grouplens.org can be found here: http://grouplens.org/datasets/movielens/latest/. This website has two datasets, a small dataset containing a list of 9,000 movies and a full dataset containing 45,000 movies. The smaller dataset has been used for this project. This dataset contains

1. Movie title
2. Genres
3. Link to IMDB (IMDB Identifier)

Using the link to IMDB identifier, additional data about the movie has been obtained from imdb.com and boxofficemojo.com websites. From the IMDB website, the following information about a movie has been extracted

1. The top 4 actors of the movie and their gender
2. The number of movies the actors have acted in till the year in which the movie was released
3. The director of the movie
4. The number of movies the director has directed till the year in which the movie was released
5. Release date
6. Budget of the movie
7. Production studio

From www.boxofficemojo.com the movie 4th weekend gross-to-date(GTD) box office information has been collected.

2.2. Dataset Characteristics

The movie dataset from grouplens has 9,000 movies, however, **complete data is available only for 1,000 movies**. Most of the movies that have been chosen have been made after the year 1980.

The following chart shows the distribution of the movies chosen by release year.

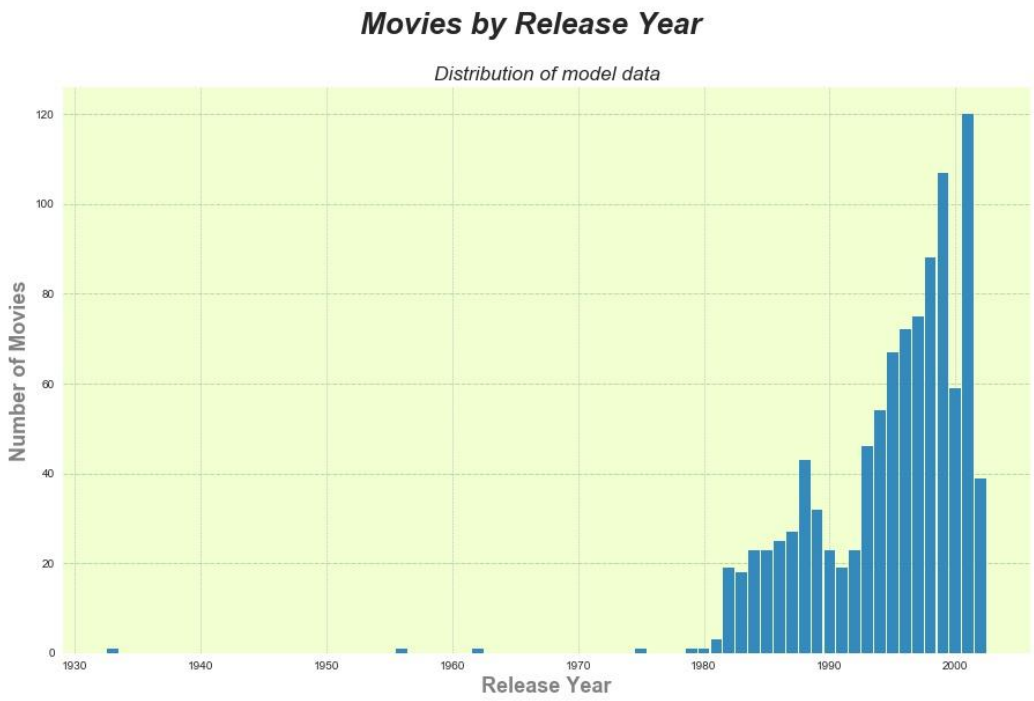

FIGURE 3 MOVIES BY RELEASE YEAR

From the above chart we see that most of the movies that have been chosen to build this mode are the ones made after 1980 and the later years have more movies. This could be an anomaly of the data that has been chosen. I tried to find the total number of movies that been made by year, while there is no definitive source for this information, there are a couple of sites which has some information. As per https://www.the-numbers.com/market/ the number of movies made from 1995 have more or less remained the same. As per http://members.chello.nl/~a.degreef/Filmnummers.html which includes movies in other languages as well, the number of movies being made hit its peak in the 1930s - 1940s. After that it starts to decline till 1970's and has been on an upward trend from there on.

The following chart shows the distribution of the movies by genre. Note that the sum of all genres will be greater than the sum of the total number of movies because some movies have been tagged with more than one genre.

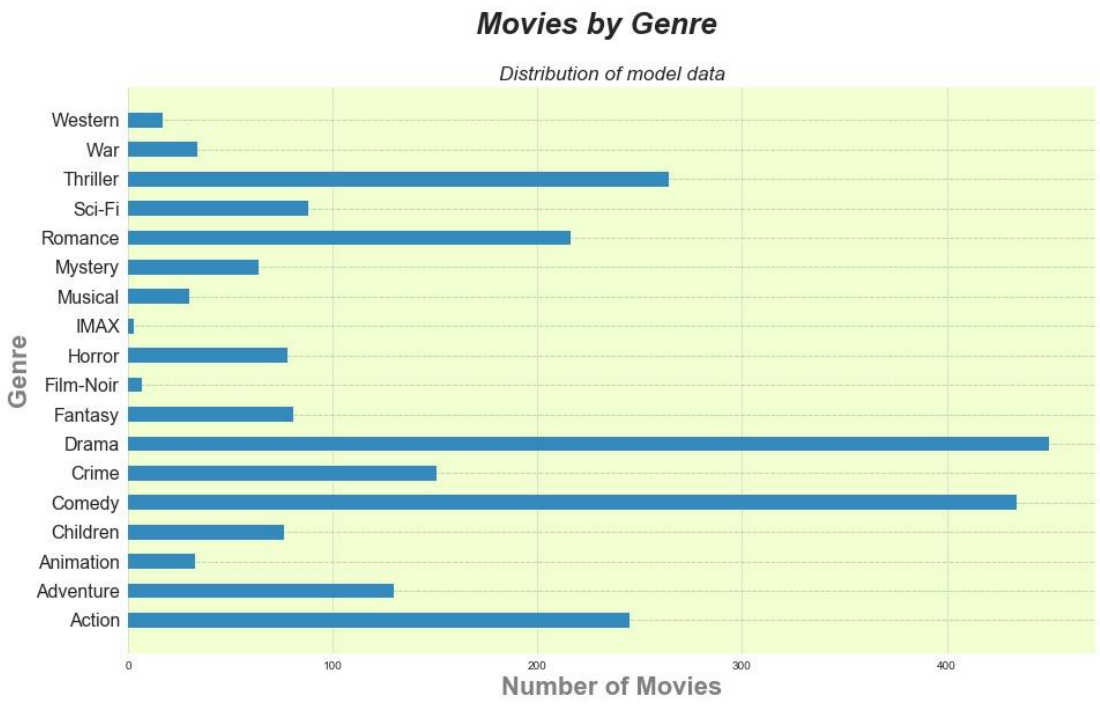

FIGURE 4 MOVIES BY GENRE

From the above chart we see that the top most genre is Drama closely followed by Comedy. The other top genre of movies are Thrillers, Action and Romance.

The following chart shows the distribution of the difference between the gross-to-date (GTD) box office collection and the movie production budget.

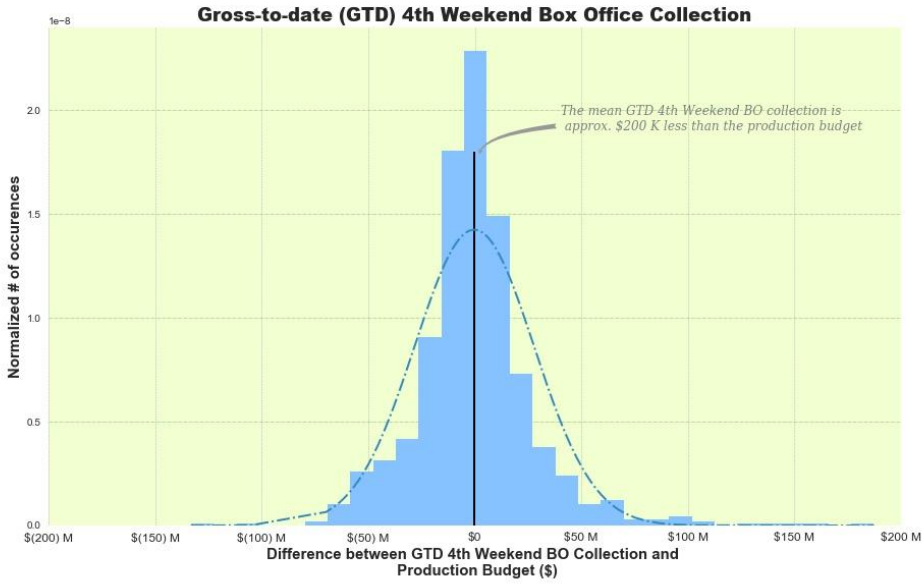

FIGURE 5 GTD 4TH WEEKEND BOX OFFICE COLLECTION

The distribution roughly follows a Normal distribution and, on an average, the movies GTD 4th weekend collection is approximately $200 K less than the production budget. At the extremes, there are few movies that have made as much as $250 Million more than the production budget and there are few that are short by $150 M compared to the production budget.

3. SOLUTION DETAILS

3.1. DATA CLEANING / WRANGLING

Following are the high-level steps in collecting and cleaning the data

1. Start with the data from grouplens.org, exclude documentaries and TV series from the data leaving with only movies
2. Split the genres in to separate columns. Essentially, create dummy columns for each genre
3. Split the movie data into smaller chunks and run the data collection / cleansing on each chunk. The reason for this is that for each movie, there would be multiple hits on imdb.com and other processing leading to a large number of consecutive hits on imdb.com and a long processing time for each movie. Splitting it into smaller chunks makes it easier to manage by limiting the total number of hits and by providing the ability to rerun the process without the need to run for the whole dataset
4. Use the defined function **`get_cast_mv_details_from_mvie(movieId)`** to get the movie details from IMDB using IMDbPY package; Below are the details of this function

```
Input:
1. Movie Id (String)
Output:
1. A dictionary containing details of the movie

Given a movie id, this function checks if the movie id is a valid movie id
and return the details of the movie if the id is valid.
```

```
A movie id is decided to be valid if the 'kind' is movie and 'English' is
one of the movie languages and 'USA' or 'UK' is one of the countries in
which the movie was released. Basically, we would like to include only
movies (not TV Shows or documentaries) that have been made in English
language and has been released in USA to build the data.
If a movie is found to be valid, this function returns the following as a
dictionary: kind, year of release, plot of the movie, production studio
and the score of the first 4 actors in the movie and the score of the
director.
An actor's score is just the total number of movies the actor has acted in
till the year in which the movie was released. Similarly, the score of the
director is the total number of moves the director has directed till the
year in which the movie was released.
```

5. Get the movie budget and release date from scraping www.imdb.com using Beautiful Soup package.
6. Similarly, get the 4th weekend gross-to-date box office (BO) collection for each movie from www.boxofficemojo.com. The function **`get_movie_4thWkndBo_BoMojo(m_title)`** performs this step. The challenge in this is that unlike IMDB, www.boxofficemojo.com does not have a unique identifier for a movie, so some of the movie BO details had to be manually fixed, especially for movies that were remade and were sequels.
7. Join all the datasets together and remove the ones that do not have data; there are multiple points where the data may not be present
    a. Movie BO data – There are many movies for which the BO data was not found
    b. Actor / Director Score – For some of the actors and directors, the scores could not be found
8. Drop the movies where data is missing
9. Some of the numeric information like movie budget was returned as string from the function; those were converted to numeric.
10. Release date was not found for some of the movies; however the release year was available; so instead of throwing this data away, the function **`fix_release_dates(date_string,year,title)`** fixes these dates by picking a random month and 15th as the day for these movies where only the release year is known

3.2. Exploratory Data Analysis

Inferential Statistics

Are there variables that are particularly significant in terms of explaining the answer to your project question?

Section 2.1. gives details on the variables / columns contained in the provided dataset. The important variables that are

1. Actor's score
2. Director's score
3. Budget
4. Release month / season

Are there strong correlations between pairs of independent variables, or between an independent and a dependent variable?

Intuitively, the main reason people go to watch a movie is either its actors or its director. Below chart shows the correlation between the lead actor's score and the measure of the success. The actors score is simply the total

number of movies the actor has acted till date. Similarly, the director's score is the sum of movies the director has directed till date.

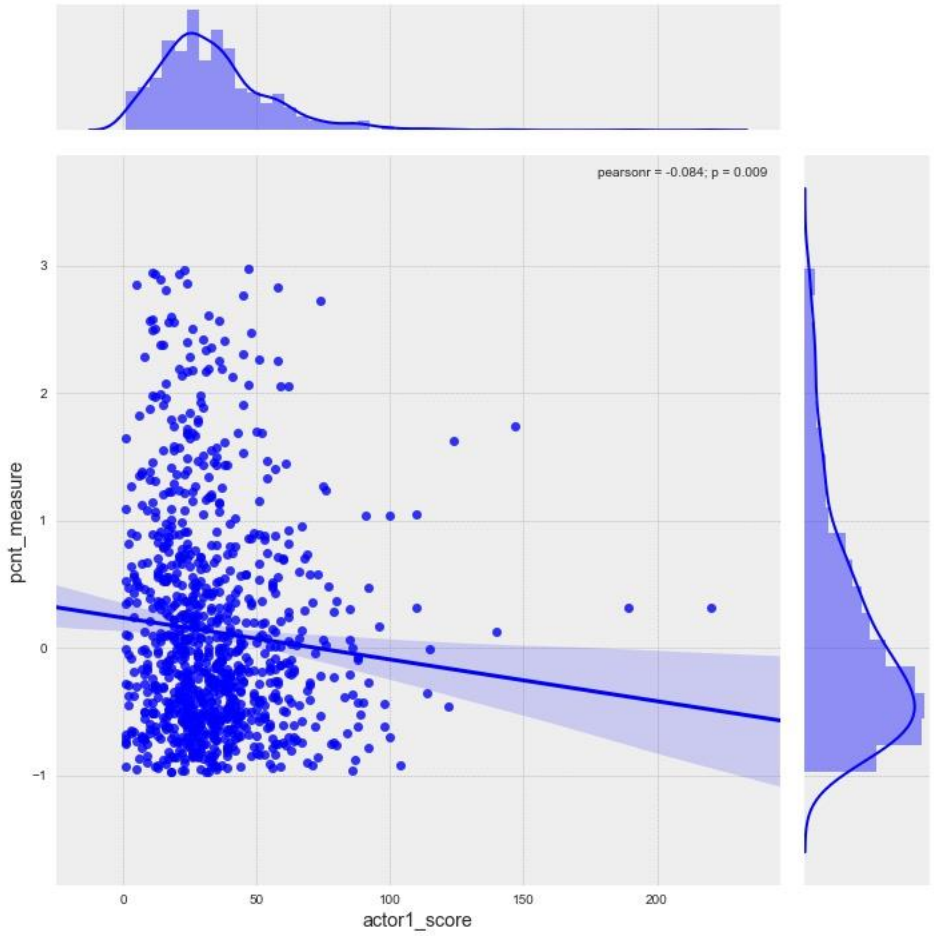

FIGURE 6 LEAD ACTOR - MOVIE SUCCESS MEASURE CORRELATION

From the chart, it seems that there is a slight negative correlation between the lead actor's score and the movie success. It could be that actors past their prime may not be as successful as upcoming actors or actors who are in the prime of their career. From the distribution on the top, one can see that the distribution of the actors score follows a skewed normal distribution with most scores between 25 to 35, i.e. they have acted in 25 to 35 movies. Similarly, below chart shows the correlation between movie success factor and the director's score.

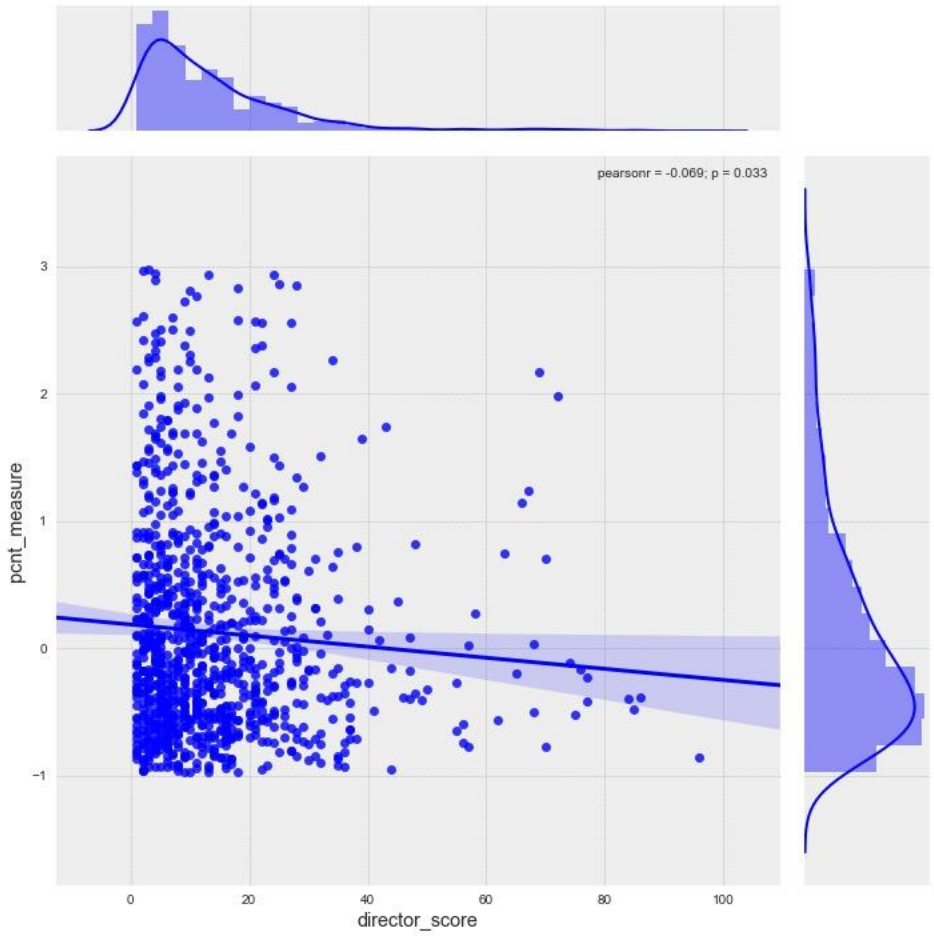

FIGURE 7 DIRECTOR SCORE - MOVIE SUCCESS MEASURE CORRELATION

The correlation between the director's score and the movie's success follow a similar patter as the correlation between lead actor's score and the movie success.

Another important factor is the month in which the movie is released. Generally, movies can be grouped by the season they are released, with movies being released with the flavor of the season. For example, holiday themed movies are released around December, romantic movies are released around Valentines Day and horror movies wait for a Friday the 13th. The below chart shows violin plots of movies grouped by the release month.

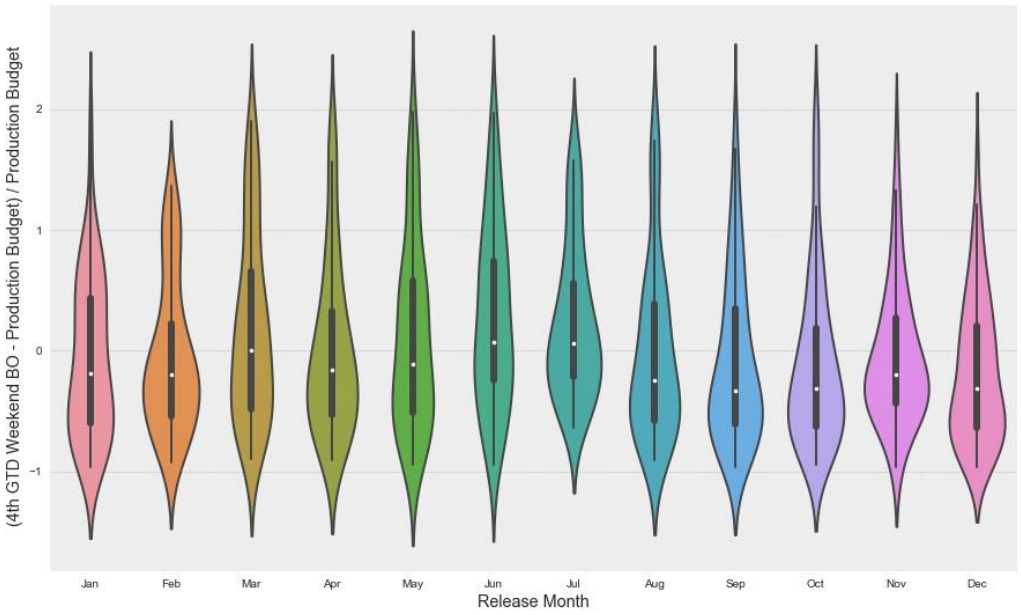

FIGURE 8 VIOLIN PLOTS OF MOVIES GROUPED BY RELEASE MONTH

The clear result from the above chart is that movies that are released during summer are more successful than any other seasons. On an average, by the 4th weekend, movies that are released during Jun and Jul would have collected more than the production budget. The next best season for movies seem to be Spring season and the worst time for movies is September. Surprisingly, movies that are released during the holiday season fare no better than any other typical month.

3.3. DATA VISUALIZATION

Another factor that has changed over the years could be the number lead roles offered to women actors. With the amount of focus being given to female empowerment, it is natural to think that the number of lead roles taken up by women would have gone up. Below chart shows the % of roles taken up by women actors.

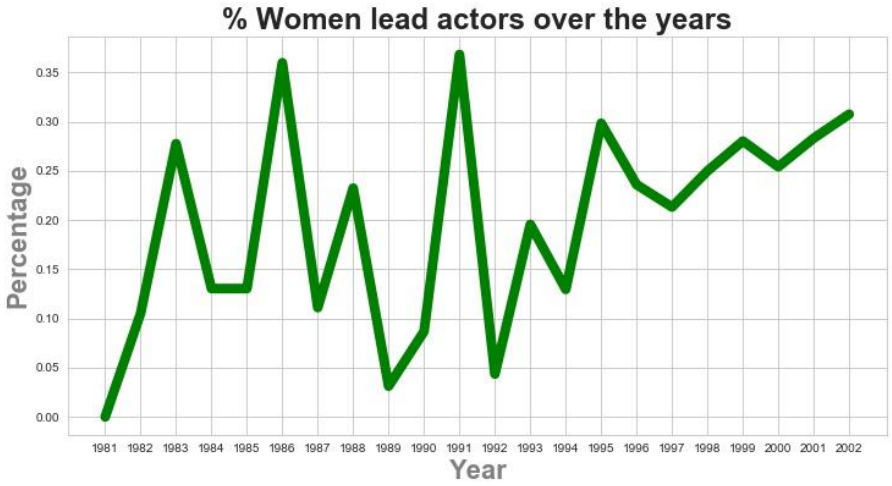

FIGURE 9 WOMEN LEAD ACTORS OVER YEARS

There is a lot of variation in % of lead roles given to women actors between the years 1980 and 2000, however, it has never gone above 35 %.

Let us look at the overall distribution of the 4th weekend BO collection for movies with Male and Female lead roles.

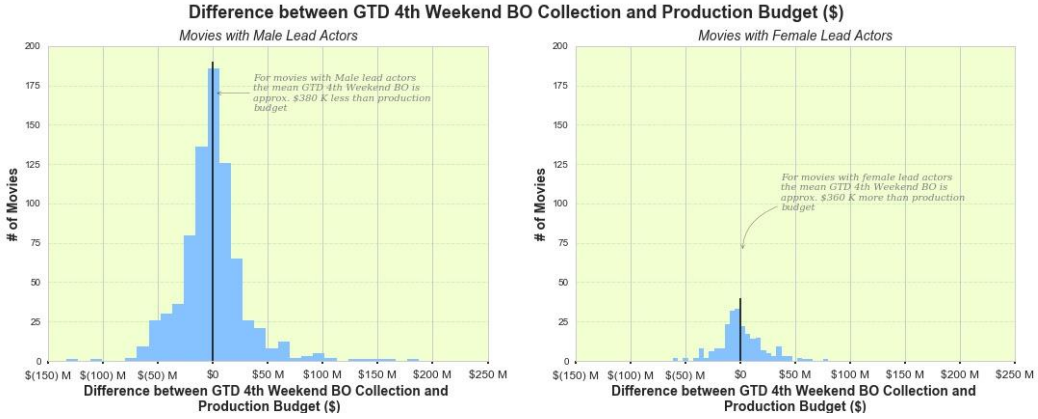

FIGURE 10 FEMALE - MALE LEAD ACTORS - MOVIE SUCCESS FACTOR DISTRIBUTION

The above charts show the difference between the gross-to-date (GTD) 4th weekend box office collection and the movie production budget. The chart on the left shows the chart for movies with male lead actors and the one on the right is for movies with female lead actors. The above charts confirms what everyone suspects that there is huge gender gap in the movie industry with most lead characters being played by male. The 4th weekend GTD box

office collection for movies with female lead actors is $ 560 K more than the production budget, whereas, the same for a movie with a male lead actor is $1.5 Million less than the production budget.

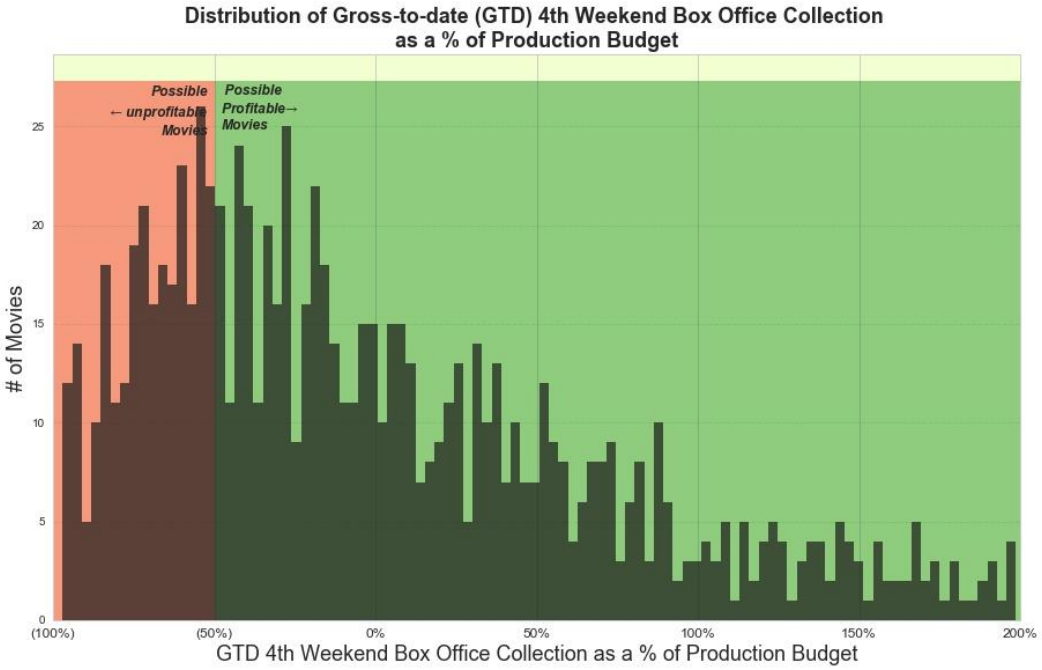

FIGURE 11 4TH WEEKEND GTD BOX-OFFICE DISTRIBUTION

The above chart shows the distribution of the percentage measure (Difference between 4th weekend box office collection as a percentage of the movie production budget). The objective of this is to find how many movies are probably successful and how many are not. There are lot of expenses that go into a movie and there is no clear information if a movie has made profit or not. The major categories of expenses are production costs, distribution costs and marketing costs. The revenue for a movie comes from many sources, the broad categories of revenue streams are box office collection, media streaming, tv rights and merchandising. Based on various materials that i could find on the internet, i have considered an approximation that by the fourth weekend, if the box office collection is at least 50% of the production budget, it is a fair indicator of movie success

Below are some pages that have some details / discussion on accounting and financial success of movies.

https://www.quora.com/How-much-money-as-a-percentage-of-its-budget-does-a-Hollywood-movie-need-to-gross-to-actually-make-a-profit

https://movies.stackexchange.com/questions/12882/how-much-of-a-films-revenue-comes-from-its-cinema-release

https://en.wikipedia.org/wiki/Hollywood_accounting

3.4. FEATURE SELECTION

As discussed in section 2.1, there are many features to choose from. Within the constraints of time and processing cost for this project, below are the features that has been used.

1. The top 4 actors' scores
2. Actors gender
3. The directors' scores
4. Movie genre
5. Release Month

A total of 46 features have been used in this model.

There are other potential features that can be used to make this model more effective.

3.5. Quality Metric

The model's objective will be to predict the following metric about a given movie:

$$\frac{4th\ Weekend\ GTD\ Box\ Office\ Collection - Movie\ Production\ Budget}{Movie\ Production\ Budget}$$

The lowest value possible for this metric will be -1 or negative 100% when the 4th Weekend GTD box-office collection is $ 0. The maximum possible value theoretically is infinity.

The quality metric used to judge a prediction model's accuracy is the residual sum of squares (RSS). It is the sum of squares of the differences between the observed responses in the dataset, and the responses predicted by the model.

Baseline Performance: The model will have to perform better than using the average of all observations as the predicted value. The RSS on the test dataset for this model where we use the average is the predicted value is 294 (1,330 with the training data). So, any model that is created for this project should have a RSS that is lesser than 294.

3.6. Machine Learning

This is a regression problem, different models using the ML algorithms have been created. For each model, each ML algorithms was run with default values, then the hyper-parameters were tuned using `GridSearchCV` function to achieve the best model.

1. Bayesian Ridge Regression
2. Random Forest Regressor
3. Support Vector Machine
4. Decision Tree Regressor
5. Extra Trees Regressor

The train-test split was taken to be 80/20.

3.6.1. ML Model Performance

The below table shows the performance of each model after tuning the hyper parameters.

Model	RSS with Test data	RSS with Training Data

Bayesian Ridge Regression	223.81	1052.82
Random Forest Regressor	211.07	798.71
Support Vector Machine	321.60	1315.26
Decision Tree Regressor	230.86	1315.26
Extra Trees Regressor	226.67	897.88

FIGURE 12 ML MODEL PERFORMANCE

While most of the models offered a reasonable improvement of accuracy over the baseline performance which had a RSS of 294, the Support Vector Machine (SVM) performed worse than the baseline. The RSS for SVM is 321.6. The Random Forest Regressor had the best performance among the models with a RSS of 211.07.

The below charts show the plot of actual against predicted values for each model.

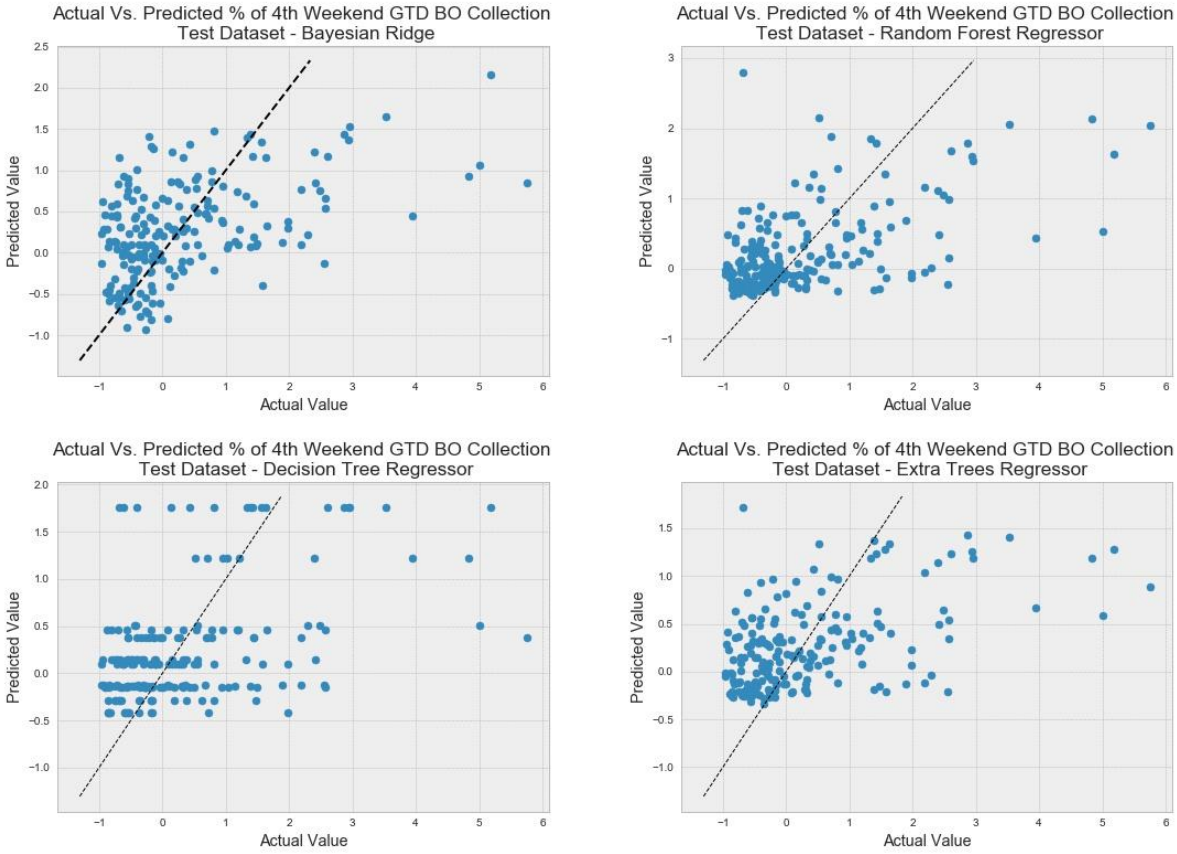

FIGURE 13 ACTUAL - PREDICTED VALUES SCATTER PLOT

From the charts, we see that Bayesian Ridge Regression seems to perform better across a wider range. Though as per RSS, the Random Forest Regression seems to work better for a narrow range (-1 < actual value < 0.5). The Decision Tree Regressor predicts the values into distinct groups.

3.6.2. Most Important Features

The generated models will help to understand the most important factors that affect the profitability of movies. The below tables show the most important features as determined by the algorithm.

Regression Model	1st Important Feature	2nd Important Feature	3rd Important Feature	4th Important Feature	5th Important Feature
Bayesian Ridge Regression	June (Month)	Drama (Genre)	Sep (Month)	Horror (Genre)	Dec (Month)
Random Forest Regressor	Movie Budget	Release Year	Drama (Genre)	4th Lead Actor Score	2nd Actor Score
Decision Tree Regressor	Movie Budget	Drama (Genre)	Release Year	Lead Actor Score	Director Score
Extra Trees Regressor	Movie Budget	Release Year	Drama (Genre)	June (Month)	Aug (Month)

FIGURE 14 FEATURE IMPORTANCE

From the above table, we see that Drama genre feature among the most important features in all the models. Other features that appear the most are the movie budget and movie release year. The most surprising factor is that lead actor's or the director's score does not seem to be the most important factor.

3.6.3. CONCLUSION

From all the above, following are the key conclusions:

1. Predicting success of a movie is a challenge, the success of a movie does not correlate strongly with any of the factors that affect a movie
2. However, there are certain features that have more importance than other in affecting the success of a movie, e.g. drama genre movies are probably more successful and movies released during summer fare better than movies released in other months

The model created here for this project is a very basic model, there are a lot of ways this can be improved. Apart from increasing the volume of observations available to run the model, some of the following additional data could be used to improve the model

- Number of theatres in which the movie was released
- Breakup of production budget, i.e. amount spent towards special effects, actors' salaries, etc.
- Experience of the production studio
- Scriptwriter's past success
- Strength of the movie script, this would require text analysis models
- Number of languages in which the movie was released
- Number of countries in which the movie was released

4. SOURCE CODE

4.1. GIT REPOSITORY

The source code for this project is maintained on Git at: https://github.com/rajeshdsar/Movie_Success_Pred_CStn2