Springboard Data Science Career Track Capstone Project II

Movie Success Prediction

By Rajesh Dharmarajan

https://github.com/rajeshdsar/Movie_Success_Pred_CStn2

Table of Contents

Movie Success Prediction

- 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
- Challenge : Could we understand which variables or factors are the most important in determining the box office success of a movie?
- Create a model that uses movie data that is available before the movie is released or even made
- Use data from sources like imdb.com and boxofficemojo.com to gather characteristics of a movie
- Potential customers for this would be anyone who has a stake or interest in movie industry

Summary of Approach

Collect

- IMDB API
- Web Scraping
- ImdbPy
- BeatifulSoup4

Explore

- Exploratory Data Analysis
- Pandas

Wrangling

- Fix Null values
- Transformation
- Pandas

Visualize

- Data Visualization
- Pyplot / Seaborn

Machine Learning

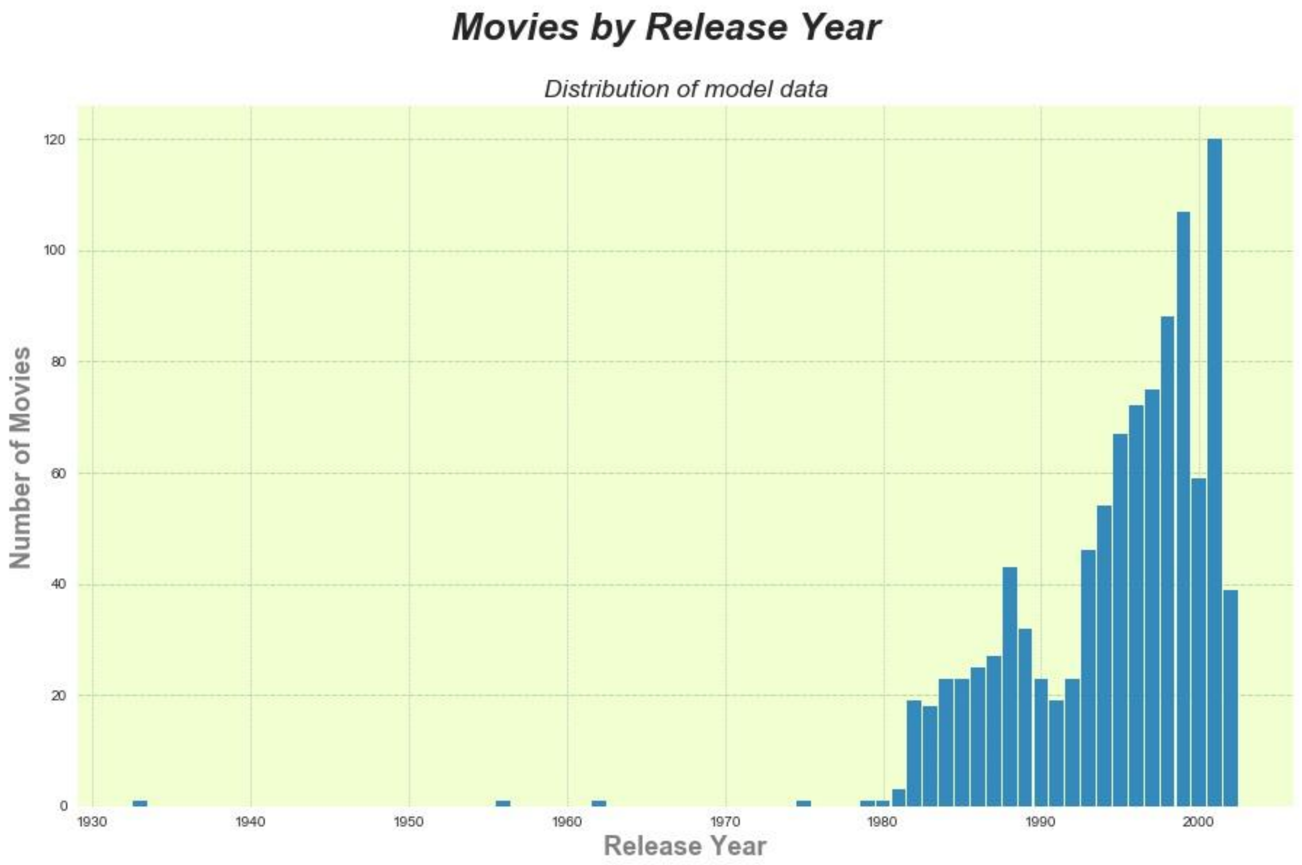
- Quality Metric
- Classification Algorithm
- Scikit-learn / Numpy

Dataset Characteristic

- Data to be used for this has been extracted from
  - www.imdb.com
  - www.boxofficemojo.com and
  - https://grouplens.org
- The list of movies from the grouplens website has been used as a starting point to get additional data from IMDB and boxofficemojo websites
- Movie dataset from grouplens has 9,000 movies, however, complete data is available only for 1,000 movies.

Dataset Characteristic

- Most of the movies that have been chosen to build this mode are the ones made after 1980



Data Wrangling

Movie Identifiers from Grouplens
- Use the 9000 movie dataset
- Exclude items like documentaries, etc.

Dummy-out categorical features
- Create dummy columns for features like genre, gender, etc.

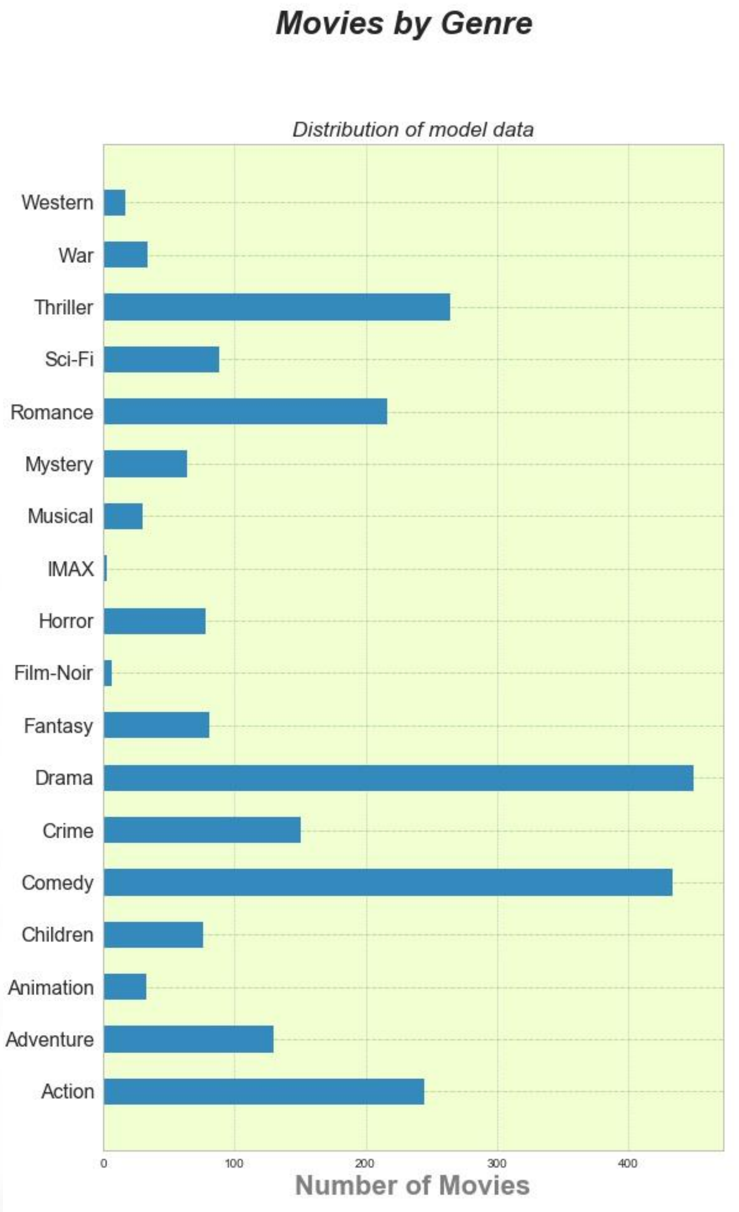
Split into smaller chunks
- Multiple hits on imdb.com, so split into smaller chunks
- Easier to manage

For each chunk use ImdbPy to get movie data
- Include only valid movies
- Get actor’s score - total number of movies the actor has acted in till the year in which the movie was released
- Get movie budget using BeautifulSoup

Join all the data chunks together
- Join the chunks together to form the original dataset

Get box-office collection data from boxofficemojo
- Use BeautifulSoup to get box office data
- No unique identifier on this website – this is a challenge

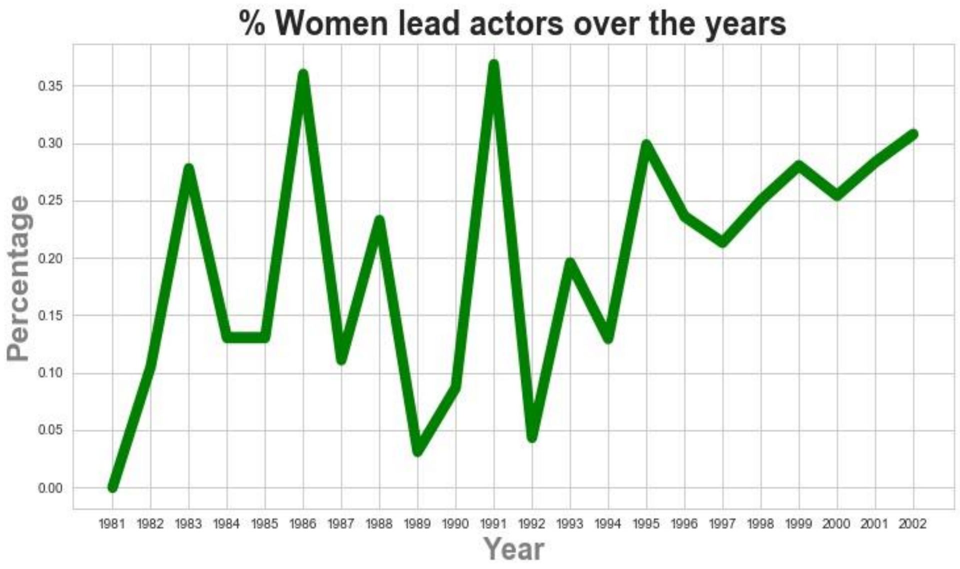
Check for missing data
- Drop records with missing data
- Convert numbers stored as text to numbers



Data Visualization

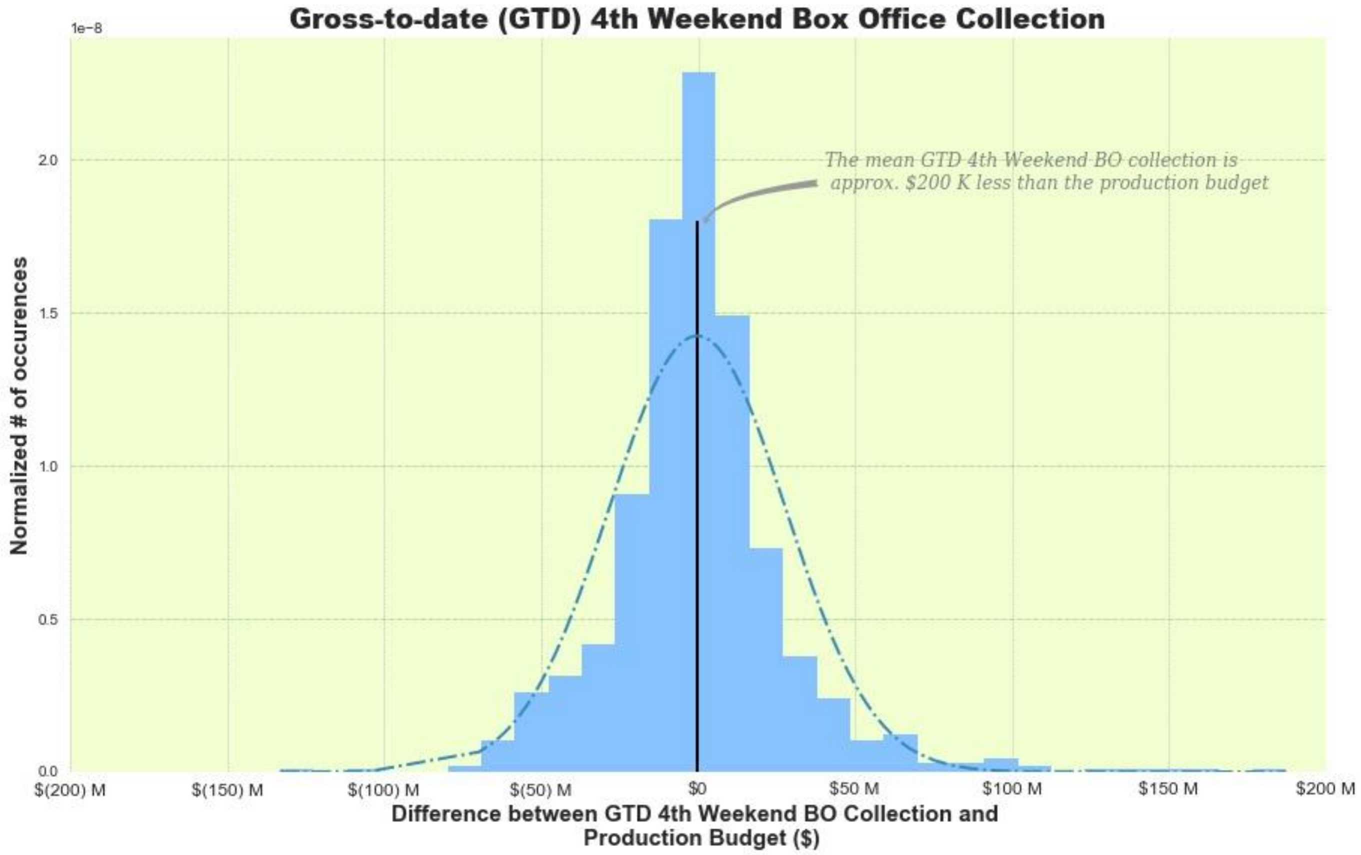
- Drama is the most popular genre, followed by Comedy and Thriller
- Each movie may be tagged with more than one genre

Data Visualization



- Though there is a lot of variation in the % of lead roles given to women actors, it has not gone above 35 % between 1980 and 2000

Data Visualization

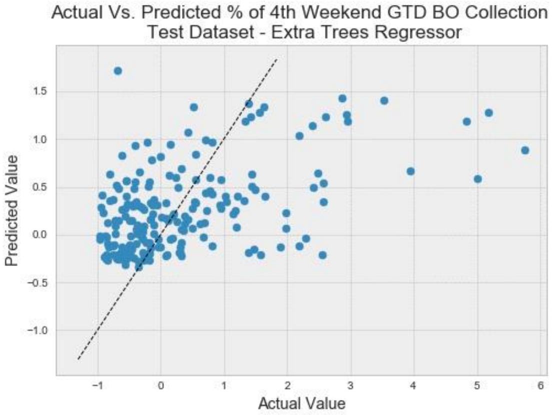


- The distribution of box-office collection roughly follows a Normal distribution
- On an average, the movies GTD 4th weekend collection is approximately $200 K less than the production budget.

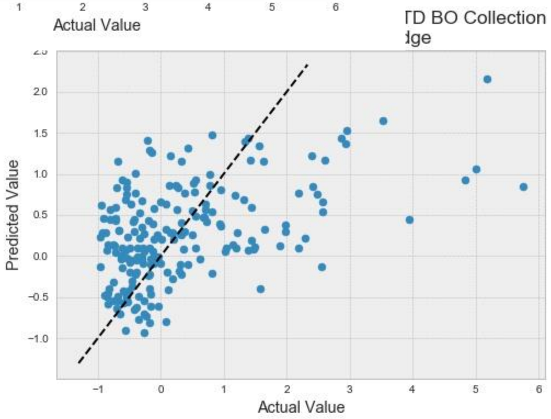
Machine Learning Models Used

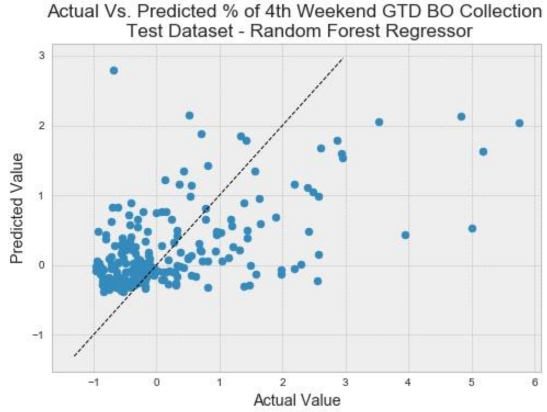
Model	RSS with Test data	RSS with Training Data
Bayesian Ridge Regression	223.81	1052.82
Random Forest Regressor	211.07	798.71
Support Vector Machine	321.60	1315.26
Decision Tree Regressor	230.86	1315.26
Extra Trees Regressor	226.67	897.88

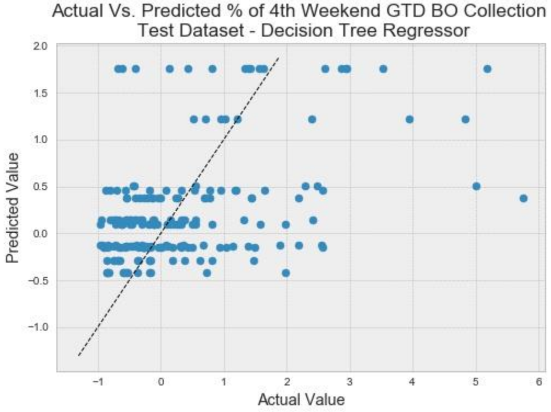
- Train-test data taken at 80/20
- Residual sum of squares (RSS) of prediction used as the metric to evaluate the models
- If the mean of the observations were used as the predicted value, the RSS with the Test data was 294
- Random Forests produced the best results; SVM had the worst performance, even lower than using just the mean of observations

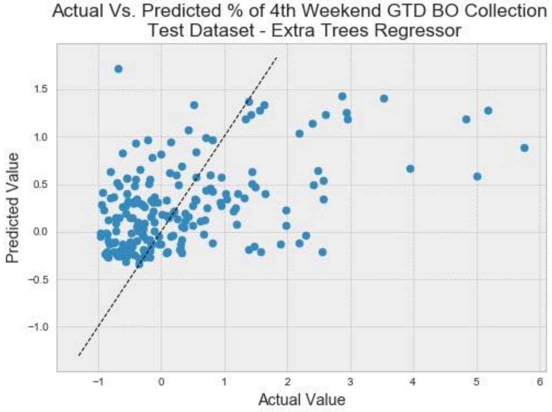


Predicted Values









- Bayesian Ridge Regression seems to perform better across a wider range.
- Though by RSS, the Random Forest Regression seems to work better, it works better only for a narrow range (-1 < actual value < 0.5)

The Most Important Features

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)

- The models help understand the most important factors that affect the profitability of movies
- Drama genre is the most important feature followed by movie budget and release year
- Lead actor’s score or director’ score are important but not the most important as per these models

Conclusion

- 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
- 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 models can be improved with volume of data and additional features like number of release theatres, scriptwriters experience, strength of the movie script, familiarity of the movie characters