Movie Performance Prediction based on Pre and Post Release Data

Rahul Handa
Department of Computer Science
UMass Amherst, MA 01002
rhanda@cs.umass.edu

Suraj Subraveti
Department of Computer Science
UMass Amherst, MA 01002
ssubraveti@cs.umass.edu

Ananya Suraj
Department of Computer Science
UMass Amherst, MA 01002
asuraj@cs.umass.edu

Abstract

We consider the problem of predicting a movie's box office revenue. This report examines the various factors that go into predicting a movie's revenue such as sentiment score of critic reviews, star popularity, genre, rating etc. The sample consists of 2217 movies that were released between 2000 and 2015. Experiments include adjusting the budget and revenue generated by a movie to account for inflation and trying a binary hit/miss style classification approach. Moreover, our experimentation also shows that some features are more important than others when it comes to predicting movie revenue.

1 Introduction

A successful movie requires an optimal combination of cast, crew, budget, plot, time of release and media attention. Even with such an optimal combination, there have been times when a movie was a complete failure in the box office. For example the recent Ghostbusters movie was a disaster despite having Chris Hemsworth, a large fan base and being a reboot of a popular franchise from the 90s. The movie was required to make close to $400 million dollars to break even from production and distribution costs. However, it made around $200 million which was approximately close to the filming budget itself. Though there might be exceptional cases and factors involved in the amount of revenue generated by a movie, the aim of this project was to analyze and determine these factors and their interaction with each other using machine learning.

Our main focus is to predict the revenue generated by movies from information that already exists. For this project we decided to limit the locale of audio track to English, but we have also considered the revenue generated by some movies that were released in multiple languages including English. In order to predict revenues, we looked at data that is available before a movie is even made, like the popularity of the actors involved in the movie, the budget, the kind of movie it's going to be (comedy, horror, action etc) and the time of release (political environments and economic factors can be a major influence too). We examine the importance of the above mentioned features with respect to the revenue a movie generates. For example, a superhero action movie with a big budget and a popular actor playing the lead role has almost proven to be a formula for success.

We also considered data that is available after a movie's release. The initial reviews by prominent critics during the week of release could have a high impact on the movie's success as this would directly affect public opinion about the movie. Further, the rating a movie is assigned (PG-13, Rated R, Rated G etc) will put a constraint on the kind of audience that is allowed to view the movie

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

and hence impacts on the ticket sales. The popularity of a movie is generally showcased by scores assigned to it on popular websites like IMDb and RottenTomatoes. This would also have an impact on the revenue as a lot of people use these websites for movie recommendations and the more popular a movie is, the more revenue it would generate.

2 Related Work

Many researchers have developed and analyzed models on movie revenue data and the factors that affect it. In 1983, Litman tried to predict the revenue of movies by using multiple regression models. In this study, features such as the genre of the film, the Motion Picture Association of America's rating (G, PG etc), the budget of the movie and the release date were taken into account. The result of the study shows that the following optimal combination in a movie was a sure shot towards success: High budget, PG or PG- 13 ratings, science fiction genre and a Christmas time release [1]. Litman et al (1989)[2] and Litman et al (1998) have both analyzed models on the role of the critic in the film industry. Their studies suggests that critics are integral influencers for box office revenues[3].

Einav (2001), in his study finds that the successful movies are usually released during the holiday seasons or vacations i.e. Christmas and during the summer. He concludes that movie ratings influence the box office revenues majorly as this affects the range of audience eligible to watch a particular movie is very important. Movies with an MPAA rating of restricted (rated R) do not seem to do very well at the box office in comparison to movies with other ratings[4]. Using a linear regression model, Ravid(1999), shows that PG and G rated movies generally do better in the box office[5]. In a study conducted by Neelamegham and Chinatagunta (1999), a Bayesian model was used which hypothesized that internationally, thrillers are more popular than romance movies[6].

In 2011, Neil Terry, Michael Butler and De'Arno De'Armond conducted research and analyzed the impact of film critics, genre, rating and awards. By using a multiple regression model, each feature involved in a movie's success was analyzed. Most of the data was collected from the Rotten Tomatoes website and the study was conducted on 505 samples of movies released between 2001 and 2003. The results of this study concluded that an increase of 10% in critic approval improves the revenue generated by around $7 million dollars. Further, movies with an R rating were found to lose more than $12.5 million dollars[7].

In the study conducted by R Parimi et al, movies were categorized into two classes based on their box office revenues (success or failure) and a binary classification was adopted for predictions. Some other studies also considered a multi-class classification problem and movies were categorized into several categories[8].

Features were categorized into "who" - cast and their popularity along with social network analysis, "what" - genre, rating and plot of the movie and "when" - release dates and average annual profits in the research conducted by Michael T. Lash et al. By using a dataset that consisted of 2506 movies, classification algorithms such as Logistic Regression and Random Forest Classifier were trained on the dataset. An accuracy of approximately 70% was achieved for binary classification of the data based on the return on investment. The star power of a movie seemed to affect the results more than any other feature[9].

In a study conducted by Mahesh Joshi et al (2010), sentiment analysis was used to generate a score for critic reviews. Linear regression was used on text and non-text (meta) features to directly predict gross revenue aggregated over the opening weekend, and the same averaged per screen. The conclusion was that text features from pre-release reviews can indeed improve over already existing prediction models to predict movie revenues.[10]

3 Dataset

We studied the different datasets and debated upon what features we want to work upon and whether we could get/create new features from the existing ones that would add any value to our dataset. We found that none of the existing datasets had what we were looking for and hence decided to collect our own data.

3.1 Data Collection

We collected data from multiple sources like IMDb, Rotten Tomatoes and themoviedb.org (TMDb) and merged it. The method of accessing data from each of these websites varies from an API (TMDb) to accessing JSON data and data from public web pages for IMDb.

We used Python to extract, parse and clean the data that we retrieved.To get a more comprehensive dataset, our system employs both scripts that interact with APIs and a web scraper to retrieve and parse HTML data from web pages.

We obtained a list of movies from the MovieLens[11] dataset. There was a list of 30,000 movies with their genres, IMDb and TMDb IDs listed in a comma separated value(csv) file. We applied filters to the list of movies and obtained additional information from the Web.

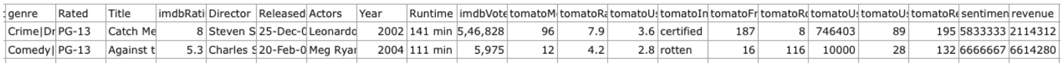
From IMDb, we obtained for each movie: movie title, IMDb rating, IMDb votes, list of genres, language, runtime, country of release, year of release, abridged list of cast, and abridged list of directors. From Rotten Tomatoes, we obtained Tomato Consensus (critic reviews), user rating, MPAA rating, studio and audience scores. From TMDb we retrieved information about budget, box office revenue and star popularity for movies from each of the above mentioned sources.

3.2 Data Cleaning

We use Pandas[12] to read and process all the data. It offers data structures and operations for manipulating numerical tables and time series. We utilize Pandas.DataFrame for data manipulation with integrated indexing. We can also read and write data between in-memory data structures and different file formats.

We removed rows that had missing feature information (for instance, the budget and revenue for some less popular movies were missing). We also collected movies that released after 2000(because 2000-2015 is a big enough timeframe and has ongoing trends in the movies), movies that had a budget <$1000 and earned less than $1000 at the box office.

After all the cleaning up the data is stored in a dataframe which in turn is saved as a csv file and looks like this.



genre	Rated	Title	imdbRati	Director	Released	Actors	Year	Runtime	imdbVote	tomatoM	tomatoRa	tomatoUs	tomatoIn	tomatoFr	tomatoRo	tomatoUs	tomatoUs	tomatoRe	sentimen	revenue
Crime|Dr	PG-13	Catch Me	8	Steven S	25-Dec-0	Leonardo	2002	141 min	5,46,828	96	7.9	3.6	certified	187	8	746403	89	195	5833333	2114312
Comedy|	PG-13	Against t	5.3	Charles S	20-Feb-0	Meg Ryan	2004	111 min	5,975	12	4.2	2.8	rotten	16	116	10000	28	132	6666667	6614280

4 Proposed Solution

4.1 Data Preprocessing

Machine learning algorithms learn from data. We decided what kind of dataset we wanted to create and collected the relevant information from the mentioned sources. But even when we have good data, we need to make sure that it is in a useful scale and in a format that can be understood by the system. Some common data preprocessing steps are formatting,cleaning and sampling . After this , we transform the data in a number of ways, viz.,

- Formatting : The data contains attributes with a mixture of scales for various quantities such as dollars, minutes and number of votes. Normalize these values
- Decomposition :Features that represent a complex concept may be more useful when split up , for example, a date can be used to extract the year and release week of the movie.
- Aggregation : Adding up the popularity value of different actors and the director of the movie. Using these techniques, we handled different features in different ways which are described below.

4.1.1 Features

We started with a dataset of 42 columns which contained some values which could not be used directly for learning like Title, plot synopsis,review consensus,etc. We used feature selection techniques to get meaningful values from them . Feature selection also helps in simplification of models to make them easier to interpret by users,shortening training times and help reduce over fitting.

Our final input feature vector contains 20 features and 2217 data cases. Versions of some features have been used in previous studies of movie revenues and profits. To our knowledge, all of these features have not been used together or at least not in the way that we have computed them.We can view these features as a pandas dataframe and can convert everything to numpy arrays for experimentation and using the data with scikit[13].The different features are preprocessed in the given ways:

1. The sentiment polarity of the critic reviews consensus which we retrieved from Rotten Tomatoes was calculated using TextBlob[14], a Python library for processing textual data. It implements a sentiment analysis algorithm based on patterns of words in a sentence to assign polarities to the input. This allowed us to calculate whether the reviews were positive or negative in order to further analyze their impact on the movie's box office revenue.

2. The cast popularity of a particular movie was calculated by taking the mean of the popularity scores of the principal actors in the cast list of that movie. The popularity scores for each actor are based on the number of ratings their previous movies received, number of followers on their facebook page and the total followers on twitter .The number of watched list additions on TMDb is also counted. This was used to analyze the impact the cast of the movie can have on its success in the Box Office.

3. The production budget of the film (in millions of dollars).The budget values were collected from TMDb and were adjusted for inflation through the CPI[15] values taken from Bureau of Labor Statistics.We adjusted the budget for each movie according to the year it came out in.

4. The country in which the movie was first released: We divided the countries class into 3 values namely 0 for USA, 1 for UK and 2 for other countries. We wanted to check this feature since some of the UK movies had large collections at the box office but ended up removing this feature during feature selection

5. Tomato certified status from rotten tomatoes : The movie will be rotten , fresh or certified, similar strategy like countries was followed .We used label encoding to transform it as 0 for rotten , 1 for fresh and 2 for certified. We can use inverse_transform to get the class values back from this data.

6. The Motion Picture Association of America (MPAA) rating of the film, one of the ratings G (general audiences), PG (parental guidance suggested), PG-13 (possibly unsuitable for children less than 13 years of age), R (children not admitted unless accompanied by an adult), NC-17 (no one under 17 admitted), and U (unrated). The ratings were encoded into 9 values using label encoder .

7. The gross revenues (in millions of dollars) for the film's general release. Again used the data from CPI to normalize this variable.

8. The IMDb rating of the movie : We aggregate the IMDdb rating and number of voters and normalize the values since the popular movies will have a much higher number of voters compared to indie and older movies.

9. The list of genres a movie belongs to (Action, Thriller, Comedy etc). We found 19 such genres and combined it into a single feature vector by using one-hot feature encoding.We checked the distribution of the movies in each genre to visualize the data.

Other features such as the runtime of the movie in minutes, the IMDb votes received by that movie from users, rating from Rotten Tomatoes, Tomato user reviews (reviews given by users of Rotten Tomatoes), the year of release of the movie and the number of IMDb votes were also considered.

4.2 Regressors

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

Since this problem involves prediction of a real value, the most intuitive approach to make predictions of revenue is to use regression techniques. We also use a binary classification approach to predict whether a movie's revenue is above or below the average revenue of all the movies.

4.3 Regression Models

Regression models involve the following variables:

- The unknown parameters, denoted as β, which may represent a scalar or a vector.
- The independent variables, X.
- The dependent variable, Y.

A regression model relates Y to a function of X and β.

$$Y \approx f(X, \beta)$$

We use Ridge Regression , Support Vector Regression, Lasso and Decision Trees with Adaboost to learn from our data.

4.3.1 Ridge Regression

Ridge Regression alleviates multicollinearity amongst regression predictor variables in a model. When features used in a regression are highly correlated, the regression coefficient of any one feature depends on which other features are included in the model, and which ones are left out. So a feature does not have any inherent effect of that feature on the response variable but only a slight effect when correlated features are included. Ridge regression adds a small bias factor to the features in order to alleviate this problem. In ridge regression, the weights are penalized using the square of the l2 norm:

$$||w||_2^2 = w^T w = \sum_{d=1}^{D} w_d^2$$

and

$$f(x) = \left(\sum_{d=1}^{D} w_d x_d \right) + b = xw + b$$

4.3.2 Support Vector Regression

The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. In the same way as with classification approach SVR optimizes the generalization bounds that are given for regression. SVR has a loss function that ignores errors, which are situated within the certain distance of the true value. This is called – epsilon intensive – loss function. Epsilon is a margin of tolerance (epsilon) within which we assume that no loss has occurred.

$$f_{SVR}(x) = \sum_{d=1}^{D} w_d x_d = xw$$

This is the equation for SVR where D denotes total number of dimensions, weights w and input x. We used a pipeline with SVR and RFE and passed it to GridSearchCV to perform hyperparameter optimization through Stratified K fold cross validation. This ensures that the variance of all the target values is high in each fold. After transformation, it is fit with SVR. The GridSearchCV helps to tune the number of features to be selected and the hyperparameter of the estimator i.e value of C and value of epsilon.

4.3.3 Lasso Regression

The Lasso is the name given to the regularized least squares when weights penalized using l1 norm :

$$||w||_1 = \sum_{d=1}^{D} |w_d|$$

The advantage here is that it simultaneously performs regularization and feature selection in order to enhance the prediction accuracy .

4.3.4 Decision Trees

Decision tree regression uses a decision tree as a predictive model which maps observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Two steps are involved in building a regression tree:

- Dividing the predictor space, i.e. $X_1, X_2, X_3, ...X_p$ into J distinct and non-overlapping regions, $R_1, R_2, R_3, ...R_j$
- For every observation falling into a given region R_j, make the same prediction, which is the mean values of the response values of all the training data points in R_j

The regions are constructed by minimizing the residual sum of squares(RSS), which in this case, is given by:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - y_{R_j})^2$$

where y_{R_j} is the mean response of all the observations in region R_j and y_i is the true value for the $i^t h$ observation

4.4 Boosting Decision Trees and AdaBoost

Boosting is an approach used to improve the accuracy of predictions resulting from a decision tree. The idea behind boosting is to use a bunch of weak predictors and combine them to increase the accuracy of the combined model. Most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. When they are added, they are weighted in some way that is usually related to the weak learners' accuracy. After a weak learner is added, the data are re-weighted: examples that are misclassified gain weight and examples that are classified correctly lose weight. Thus, future weak learners focus more on the examples that previous weak learners misclassified.

4.4.1 AdaBoost Algorithm

AdaBoost, or *Adaptive Boosting*, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire It can be used with many other types of learning algorithms to improve their performance. The output of the other learning algorithms is combined into a weighted sum that represents the final output of the boosted model.

AdaBoost is adaptive in the sense that subsequent weak learners are modified in favor of those instances classified wrongly by previous classifiers.

AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing (e.g., their error rate is smaller than 0.5 for binary classification), the final model can be proven to converge to a strong learner.

4.5 Performance

	Ridge	SVM	Lasso	D-trees-Ada
Scores	0.7094	0.7311	0.7101	0.7495

We see that we get better results with Adaboost with Decision trees. Some of the sample predictions we get with the AdaBoosted Decison Tree regressor(number of estimators= 100 and maximum number of splits=3(obtained by GridSearching over a range of hyperparameters)) :

	The Avengers	The Bourne Legacy	Underworld: Awakening	Prince Avalanche
Actual	623279547	113165635	62321039	204951
predicted	332604946	126510673	31601560	-4307757
error Percentage	46.36	-11.79	49.29	2201

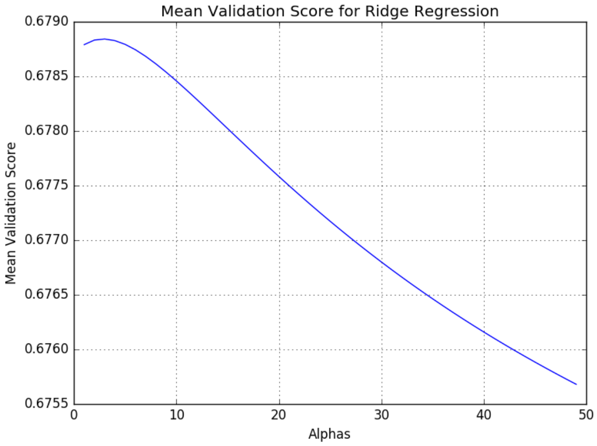
The general trend we see in all these regressors is that the big budget movies which have a sizable collection have less than 50% error rate . On the other hand the independent movies with extremely low budget and unknown actors are predicted to incur losses(negative revenue prediction). Also the SVM was able to identify some outliers in the data where the predicted values were closer to the actual real world data than the data we collected from some sources like TMDb which had old data for those movies , for eg.

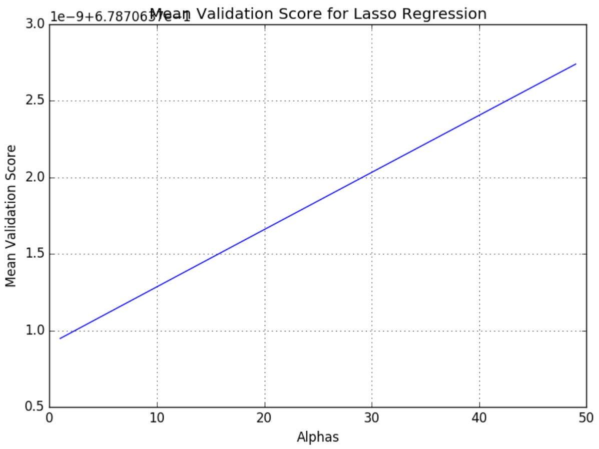
	Chalet Girl	Henry's Crime
True label	1201	100000
predicted	2830361	2271420
Real world	4811510	2200000

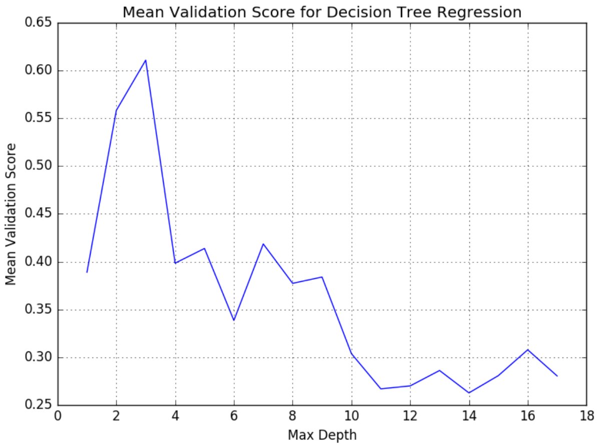
We can see how this not only predicts near real world value but also helps improve the dataset . In 60 % of the cases the predictions with error rate greater than 300% led to the discovery that the true label is in fact outdated.

4.5.1 Cross Validation

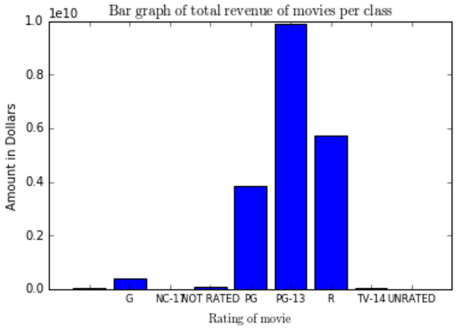
We used Gridsearch for getting the optmized hyperparameters for all the regressors and found the behavior of each regressor to be quite interesting. Here are the plots of Mean Absolute Error,which calculates the mean of absolute errors occurred while predicting on the held out data of each fold during cross validation:

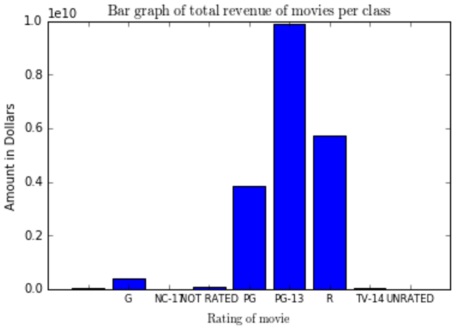






As a measure to verify our predictions numerically instead of error rates , we decided to check the total revenue predicted for each rating(R,PG,PG-13,etc) and compare it to the total ground truth revenue for each rating :





5 Experimentation

5.1 Accounting for inflation

We were predicting movie revenues in a long interval of time(15 years). Because of inflation, prices have shot up, and we thought that this could affect the accuracy of our predictions because the worth of a dollar in 2001 would be different from what it is today. As a consequence, movies released in recent years will find it easier to perform in the box office. To get all the movies on a level playing field, we decided to adjust the budgets and revenue to account for inflation. We did this by using the Consumer Price Index(CPI).

A CPI measures changes in the price level of a market basket of consumer goods and services purchased by households.

The U.S. Bureau of Labor Statistics reports the CPI on a monthly basis. Two types of CPIs are reported each time. The CPI-W measures the Consumer Price Index for Urban Wage Earners and Clerical Workers. The CPI-U is the Consumer Price Index for Urban Consumers. It accounts for 89% of the U.S. population and is the better representation of the general public. The CPI-W is a subset that covers 28% of the U.S. population.

We found this data and adjusted the budgets and revenues of the data we collected by using the following formula:

$$\textit{Adjusted Budget/Revenue} = \left(\frac{CPI(2015)}{CPI(\textit{Year of release of the movie})}\right)(Budget/Revenue)$$

When we ran our regression pipelines to this adjusted data, we did see a very small growth in the R^2 score, as shown in the table below.

Score	Ridge	SVM	Lasso	D-trees-Ada
default	0.7094	0.7311	0.7101	0.7495
inflation adjusted	0.7096	0.7299	0.7195	0.7533

The very small jump in accuracy can be probably explained by the fact that we already included the year of release as a feature in our original dataset.

For the same movie with different regressors :

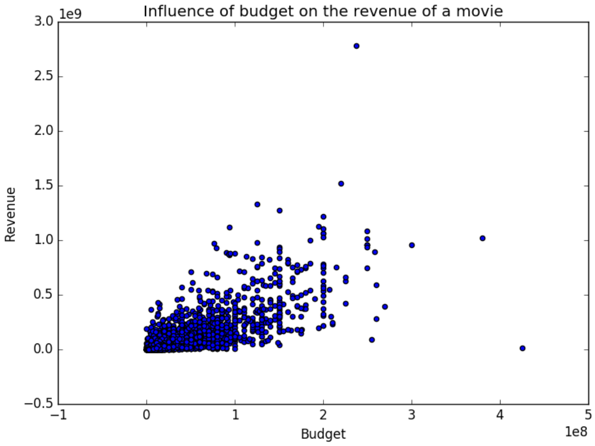
The Avengers	Ridge	SVM	Lasso	D-trees-Ada
Actual	623279547	623279547	623279547	623279547
Predicted	178282511	180724378	332604946	414984497
Error	71.39	71.00	46.63	33.4

5.2 Feature dependence on revenue

We were also curious to see what factors had a greater effect on the revenue of a movie, so we created scatter plots of certain features against the revenue to look for some patterns.

5.2.1 Budget v/s Revenue

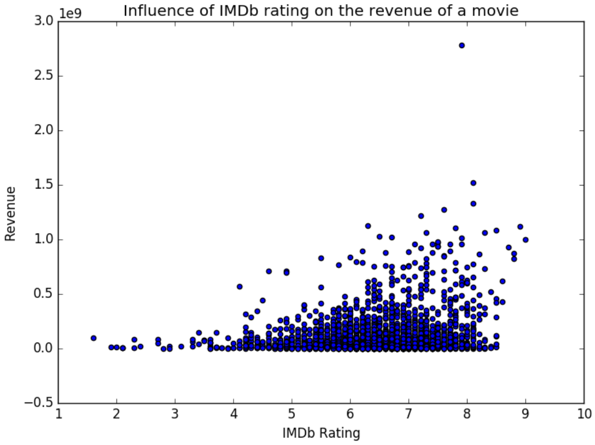
Here is the scatter plot that shows the dependence of revenue on budget of the movie:



Barring very few outliers, the scatter plot does indicate that movies with high budgets do perform better in the box office.

5.2.2 IMDb rating v/s revenue

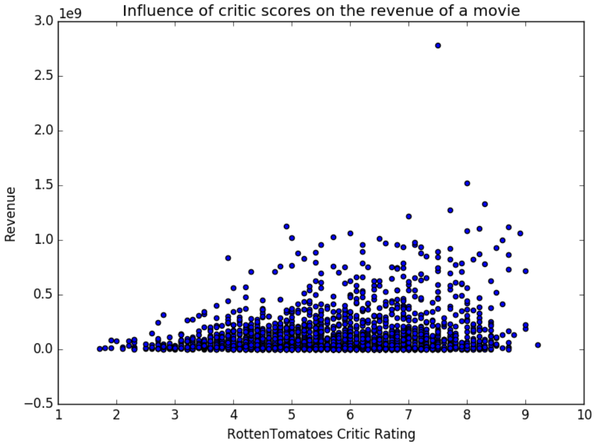
The assertion that the IMDb rating of a movie doesn't reflect it's performance in the box office is very popular. So we wanted to see if our dataset could prove/disprove this statement.



It is clear from the scatter plot that there is no straightforward relationship between IMDb rating and the box office performance. All we can assert with a certain degree of confidence is that movies with low IMDb ratings definitely do not perform well in the box office, but that doesn't guarantee record breaking earnings for movies that have a high IMDb rating.

5.2.3 RottenTomatoes Critic Rating v/s Revenue

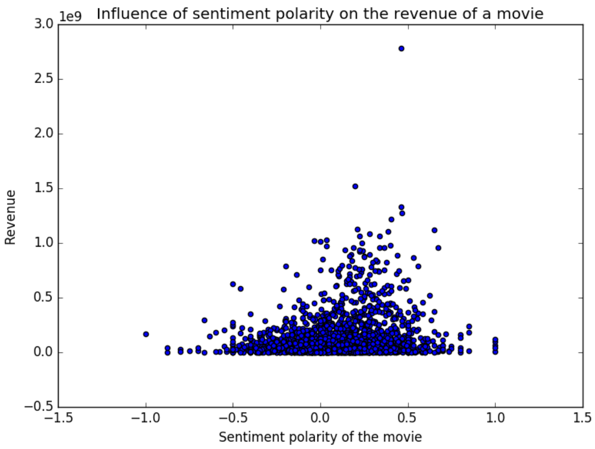
RottenTomatoes has always been the place to visit to determine the quality of a movie, because of the excellent reviews written by critics. They have a very unique rating system. Naturally, we wanted to see the influence of ratings given by expert critics on the revenue.



The results were more or less similar to what we observed with IMDb ratings. A movie that was scored badly almost always performed pretty badly in the box office, but this did not guarentee stellar performances by those movies that were scored well.

5.2.4 Sentiment Polarity v/s Revenue

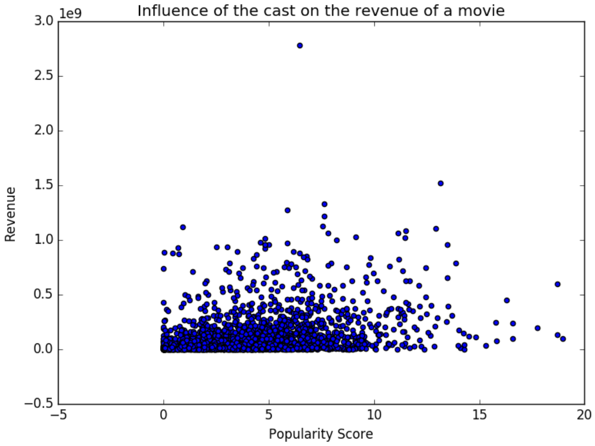
Since this was a feature we derived from the Critics Consensus from RottenTomatoes, we wanted to see how good the sentiment scores were, or what impact critic reviews would have on revenue.



The results we found were pretty interesting in that movies with high sentiment scores(i.e. very good reviews) didn't do that well in the box office. The ones with the highest revenues were the ones with decent(but not the) sentiment scores. The ones with bad sentiment scores, of course didn't do well.

5.2.5 Popularity score v/s revenue

One of the most important features we thought would influence revenue was the popularity of the actors starring in it.



There seems to be a slight co-relation between the revenue and the popularity of the starring actors. Although, there are cases in which movies with less popularity score do quite well at the box office.

5.3 Classification Approach

We performed binary classification on the data to see if the movies would be properly categorized as a success or a failure. We consider a movie to be a success if the total revenue generated by it is greater than the mean of all the revenues generated by the movies in our dataset. Otherwise, the movie is deemed to be a failure.

We considered the average revenue to be around 124 million dollars and found that 1574 movies fell into the failure category whereas 641 actually did well in the box office.

In order to experiment with these classes, we trained a number of classifiers on our data. The results look promising. Each classifier was used in a pipeline with GridSearchCV to perform hyperparameter optimization through Stratified K fold cross validation. This ensures that the variance of all the target values is high in each fold. We have also used bagging which builds several instances of a black-box estimator on random subsets of the original training set and then aggregate their individual predictions to form a final prediction. These methods are used as a way to reduce the variance of a base estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it. The results of our experiments are tabulated as follows:

Classifier	Logistic Regression	Gaussian Naive Bayes	SVM	Decision Trees	BaggedKNN
Accuracy	0.8489	0.8307	0.6965	0.8251	0.8391

6 Discussion and Conclusions

We analyzed the results obtained through our experiments which clearly show that factors such as critic reviews and star popularity are highly integral towards determining the revenue that a movie makes. We found that despite having very good reviews, some movies did not do that well in the box office because other factors would have had an equal bearing on the result too. The movies that did become hits in the box office had a pretty good sentiment score of critic reviews. High budgeted movies almost always seem to perform well in the box office especially if they have been rated PG or PG13 as this would make the movie eligible to a larger section of audience than say a movie with an R (Restricted) rating. We have also accounted for inflation in dollar value over the years and adjusted budgets and revenues generated by movies accordingly. We observed that doing this gave us an increase in the accuracy of our predictions.

One of the more interesting findings of our experiments was that since we are not giving any prior information about a movie being part of a series or some fictional universe , the revenue predicted for that particular movie is Less than the actual value even .Such movies tend to perform well in real life regardless of the critic ratings sometimes . So we could add this feature vector to our dataset and see how things work out.

References

[1] Litman, Barry R. (1983). Predicting Success of Theatrical Movies: An Empirical Study. Journal of Popular Culture, 16 (spring), 159-175.

[2] Litman, Barry R.& A. Kohl (1989). Predicting Financial Success of Motion Pictures: The 80's Experience. The Journal of Media Economics, 2 (1), 35-50.

[3] Litman, Barry R. & H. Ahn (1998). Predicting Financial Success of Motion Pictures. In B.R. Litman, The Motion Picture Mega-Industry, Allyn & Bacon Publishing, Inc.: Boston, MA.

[4] Einav, Liran (2001). Seasonality and Competition in Time: An Empirical Analysis of Release Date Decisions in the U.S. Motion Picture Industry. Working Paper, Harvard University.

[5] Ravid, S. Abraham (1999). Information, Blockbusters, and Stars: A Study of the Film Industry. Journal of Business, 72 (4), 463-492.

[6] Neelamegham, R. & P. Chinatagunta (1999). A Bayesian Model to Forecast New Product Performance in Domestic and International Markets. Marketing Science, 18 (2), 115-136

[7] Neil Terry, Michael Butler, De'Arno De'Armond (2011). The Determinants of Domestic Box Office Performance In the Motion Picture Industry. Southwestern Economic Review, 137-148.

[8] R. Parimi and D. Caragea. Pre-release Box-Office Success Prediction for Motion Pictures. Machine Learning and Data Mining in Pattern Recognition, pages 571–585, 2013.

[9] Michael T. Lash and Kang Zhao (2015). Early Predictions of Movie Success: the Who, What, and When of Profitability .

[10] Mahesh Joshi, Das, Gimpel,Noah A. Movie Reviews and Revenues: An Experiment in Text Regression Human Language Technologies: The 2010 Annual Conference of the N.A. Chapter of the ACL, 293–296

[11] http://grouplens.org/datasets/movielens/20m/

[12] pandas: a Foundational Python Library for Data Analysis and Statistics; presented at PyHPC2011

[13] API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.

[14] https://textblob.readthedocs.io/en/dev/

[15] http://www.bls.gov/cpi/tables.htm