**A Machine Learning Approach to Box-Office Profitability**

**Joseph Eklund, Kevin Johansen, David Olson, Jaryn Seidl**

CS 478, Winter 2016

Department of Computer Science

Abstract

In the film industry the main goal is to create profit-ble films. Our project takes a machine learning approach to predicting whether or not a movie will be profitable. For the purposes of this experiment, a movie is considered profitable if it earns twice as much as it cost to produce the movie. Before a movie is released, our machine learning model predicts whether its total gross income will be at least double the budget. Predictions are based on key features relating to the production of the movie including number of theaters released in, MPAA rating, and genre. The best classification accuracy we achieved was 75% using the J48 implementation of the C4.5 machine learning algorithm provided by WEKA1 Introduction

1 Introduction

It is estimated that the movie industry “will have generated 564 billion U.S. dollars in [revenue](http://www.statista.com/statistics/237769/value-of-the-us-entertainment-and-media-market/) by the end of 2014” with that estimate rising to “over 679 billion US dollars in value over the next four years” [<http://www.statista.com/topics/964/film/>]. With hundreds of millions of dollars on the line, it is in a studio’s best interest to know whether their movies will be profitable. Having an accurate prediction of movie profitability would not only help studios make more economically sound production decisions, but decisions that result in more movies that people like.

Profitability in and of itself is subjective. Many studios have different definitions of profitability, sometimes going as far as using tricky accounting methods to make a movie not profitable for tax purposes. This is the case for *Return of the Jedi*, which, despite grossing over $572 million, and a production budget of only $32.5 million, has still technically never been declared profitable by the studio [<http://www.slashfilm.com/lucasfilm-tells-darth-vader-that-return-of-the-jedi-hasnt-made-a-profit/>]. To be able to include movies ranging over the past century in our data, we decided on using a 2:1 ratio of proceeds versus production budget to define profitability. If a movie was predicted to fit this ratio, it was profitable. Proceeds were limited to worldwide box office revenue and did not include DVD/Blu-Ray/Digital sales. A Production budget was defined as the amount of money used to create the movie itself and did not include distribution or marketing.

2 Methods

2.1 Data Sources

We began our project using an open-source data scraper designed to retrieve data from HTML files pulled from BoxOfficeMojo.com. The scraper gathered values on the following features: domestic total gross income, genre, production budget, MPAA rating, distributor, release date, foreign gross income, opening week

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Definition** | **Representation** | **Example** |
| Production Budget | Total production budget in USD | Numeric | 200000000 |
| Worldwide Revenue | Total worldwide box office proceeds in USD | Numeric | 2056072722 |
| Distributor | The studio that financed the film | Nominal | Disney |
| MPAA Rating | The rating of the film by CARA (e.g., G, PG, PG13, R) | Nominal | PG-13 |
| Runtime | Total runtime of the film in minutes | Nominal | 136 |
| Number of Theaters | Number of theaters the film was in at the widest release | Nominal | 4134 |
| Release Month | Month of the year at time of release ranging from 1-12 | Nominal | 12 |
| Genre | Narrative categorization of the film | Nominal | Sci-Fi |
| Director | The director of the film | Nominal | J. J. Abrams |
| Composer | The composer of the film | Nominal | John Williams |
| Top Billed Actor | The actor who received the highest compensation | Nominal | Harrison Ford |
| Marketing Budget | Total marketing budget in USD | Numeric | 100000000 |
| Distribution Budget | Total distribution budget in USD | Numeric | 50000000 |

end income, the number of theaters the movie was released in, and the date the movie left theaters. These features provided a great starting point for our project, and we intended to add more features including director, composer, and top-billed actors. Due to time constraints we were unable to get the scraper working with the extra features we intended to pull. We decided to forego adding the additional features and run the scraper in its initial state on the approximately 16,000 movies provided by the author of the scraper. The scraper was setup to read the list of movies and send an HTTP request to BoxOfficeMojo.com to get the required HTML for each movie. While running the scraper on a sample set of 10 movies, it was clear that many independent movies included in the list would be missing a lot of data, with some movies only having a title available. We had hoped that after running the scraper on the full list, we could filter out the instances that were missing many features.

Unfortunately, while running the scraper on the full list of movies, our connection to BoxOfficeMojo was shut down. After some investigation, we discovered that scraping and data mining are against BoxOfficeMojo’s terms of service. At this point, given our limited time, our only option was to gather data manually. We got a list of 200 movies from Wikipedia, and split them up amongst our team. Each member of our team was assigned 50 movies, and we gathered data from Google and BoxOfficeMojo.com. After some discussion on what we believed would be important in predicting profitability, and the limitation of gathering our data by hand, we arrived at our final, reduced set of features: production budget, worldwide gross income (domestic + foreign income), genre, runtime, distributor, MPAA rating, opening number of theaters, and the month of release. The above table shows our finalized features with their respective representations. The red features were dropped from the final data set and are listed for reference and future analysis.

2.2 Models

Because of the nature of the data, with most features being nominal, our initial selected machine learning model was a J48 decision tree. The selection of a model did not take into account the speed of the algorithm, as long as it completed within a reasonable amount of time. The reason for using a J48 tree was because although the problem data lent itself very well to a decision tree overall, some of our initial data was continuous. ID3 does not include the handling of continuous values, so we had to either use J48 or convert our continuous data into binned data (we ended up doing both). ID3 also does not prune, whereas J48 does. We still decided to try using ID3 after trying J48 to see if we could get a more generalized (as ID3 doesn’t use all the tools that J48 does) accuracy.

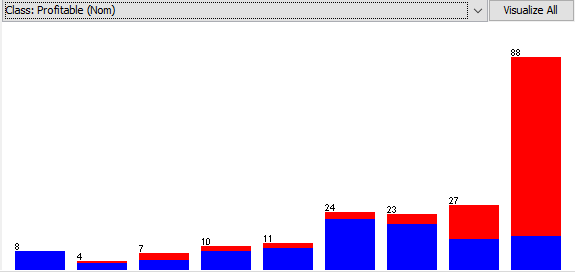
K Nearest Neighbor was another machine learning algorithm that we attempted, though the original data set would not work without being preprocessed into something that KNN could recognize. We accomplished this by doing a Principle Component Analysis on the data set, turning our data features into continuous values. See the following table for our results for each algorithm.

|  |  |  |
| --- | --- | --- |
| **Algorithm Used** | **Root Mean Squared Error** | **Classification Accuracy** |
| KNN | 0.532 | 60% |
| ID3 | 0.508 | 53.9% |
| J48 | 0.484 | 70% |

3 Results

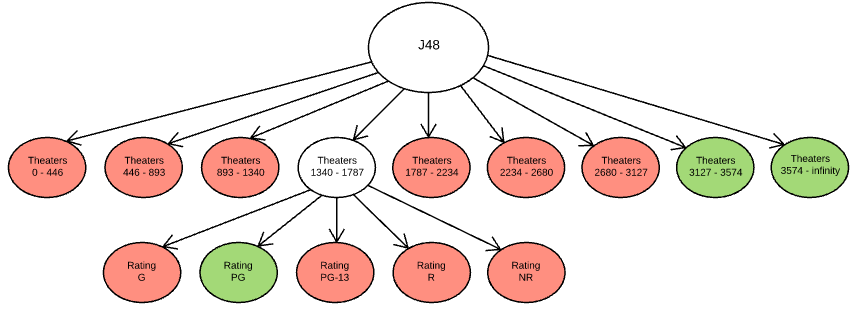
Our initial dataset contained both nominal and continuous data.  Because of this, initially a J48 decision tree was used. The J48 decision tree is capable of handling both nominal and continuous data.  It is an extension of the ID3 decision tree in that it uses the metric of information gain to decide the organization of the tree. One improvement is that it can handle the continuous values without having the user bin them.  Initially the classification accuracy was 70%. 10 fold cross validation was used for training for this model and for all following models.

With a baseline of 50% classification error, 70% is not very impressive.  Other models were attempted to see if they would yield better results.  To try other models, our initial dataset had to be transformed.  First the dataset was transformed to only contain nominal data.  Using Weka, the runtime and number of theaters were discretized into equal width bins.  Runtime was binned into 4 bins and number of theaters was binned into 9 bins.



With this new all-nominal dataset, we decided to use an ID3 tree.  The accuracy was barely better than baseline.  This is most likely due to Weka’s implementation of ID3. They do not prune to receive better results.  Also, if an instance in the test set does not exist in the training set, then it returns unclassified.  53% were classified correctly, 24% was classified incorrectly, and 23% were unclassified.

Next the data was transformed into an all-continuous dataset.  This was achieved by using

*J48 decision tree final result*

PCA.  PCA achieved two things: returned an all-continuous dataset and returned a new set of features.  This way we could try different learning models and there was the possibility of having a set of high quality features.  34 new features were found and the top 15 features were used in our models.  They all had eigenvalues over 0.7.  This feature space is larger than our initial feature space of 7, but a larger feature space was desired to improve results.  K Nearest Neighbors was used on this dataset and found decent results with k = 3:  60% accuracy with a 0.532 root mean squared error. It was an improvement on the ID3 tree but worse than the J48 tree.  Other k values were chosen but with very little impact on performance.

In Weka, there are a couple advantages that the J48 implementation has over the ID3 implementation. First, it does pruning, which gets rid of the insignificant branches.  Second, it always classifies instances.  The J48 tree was applied again but this time on the all-nominal dataset.  One problem with J48 on continuous values is that it turns continuous features into binary features by selecting some threshold to compare the value to.  If the value is greater than the threshold the value becomes 1, and if the value is less than the threshold then the value becomes 0.  So there was a greater chance of improving results if they were appropriately binned.  This provided better results:  75% classification accuracy, a 5% improvement from before. 75% still is not impressive, but it is a significant improvement to what was seen before.

**4 Conclusion and Future Work**

Using machine learning to predict box office profitability is a complex and new problem.  The feature space is hard to work with, and we only explored a part of it.  From the tests that we were able to run, we found that the number of theaters that a movie is released in ends up predicting the profitability fairly well.  We believe this is because major studios that have a lot of resources can market their product well and get the film out to thousands of theaters.  With the popularity of the movie industry, these major movies will most likely be profitable.  Other possible indicators of success could be rating or genre, but no strong conclusions were made here.  As far as machine learning models go, we found J48 to be the most complete and accurate model for this type of a problem.

There are many areas left that we could have explored given more time. A robust machine learning model for this program would include a way to automatically scrape any movie data it needs from the internet. This would allow testing of thousands more instances, instead of the couple

hundred we collected in a more laborious way. We would have also been able to get more features like the director and actor and see if those are big indicators of profitability. Also given more time, there are certainly different feature combinations that could be tested, and even different standards of profitability that could be tried, since profitability is such a subjective measure in the movie industry.

Box office profitability may also depend on other factors that aren’t so easy to quantify.  The amount of money spent on marketing is an easily quantified feature, but how it was marketed also makes a difference. There are interesting ideas about using internet activity to predict whether a movie will be profitable. People talk about the movie on Twitter and other social media sites, and many view the trailer on YouTube. We could analyze the amount of traffic involving the movie, the sentiment of the comments, and so forth. There is a lot of exploration into this type of box office profitability prediction that can be done.