# Predicting Movie Profitability Using Decision Trees

Warren Lacaba, Blessings Manatsa, Min Gyu Park

**University of Manitoba**

**COMP 4710 Project: Final Report**

**19/12/2017**

## Abstract

The movie industry is one of the most important branches of the entertainment industry, which generates a lot of revenue. The person playing a big role in this aspect is the producer as they are in charge of funding needed to produce the movie. However, producing a movie has its risks; one being that there is a chance of the movie not covering production costs. A producer relies on tools to predict profitability in movies for decision making with regards to whether or not to produce a movie project. For several years now, researchers have used different approaches to collect information that would be used as variables when predicting the success of a movie, but very few have explored using attributes directly related to a movie.

This paper focuses on using decision trees to characterize and predict movie profitability. Decision tree classifiers are relatively fast compared to other classification methods and are easily interpreted by humans. For our project, we want to see the difference between using Gini Index and Entropy for the selection of the best split point based on an attribute using an impurity function. The decision tree will be used to forecast the profitability of a movie before its production. Decision trees are models commonly used as decision support tools and its results show that the resulting model predicts whether or not a movie will be profitable with an average accuracy of 63.79%. Keeping in mind that the approach presented in this paper is not a standalone tool, it should, however, be able to round out forecasting methods such as the producer's foresight and judgment.

## Introduction

The term ‘entertainment’ refers to the entirety of all activities that generate pleasure and amusement for human beings during their free time. One major group in this industry is filmmaking and it is one of the most popular industries in North America, grossing over $10.4 billion from over 700 movies released in 2014. The movie industry is a highly dynamic industry[5]. Despite these numbers, the financial success of a movie is still very uncertain as there are examples of ‘flops’ released almost every year, hence, studios often tend to survive on few major blockbuster hits Such as “Star Wars: The Force Awakens” from 2015 that grossed 157% of its budget. A movie’s ability to earn profit depends on many factors such as the financial, social, commercial and technical factors and most of the time, these issues are not directly controlled by the producers. While researchers have tried to predict a movie’s success using various approaches, they have attempted to predict box office revenues or theater admissions. Hence investors want assurance that their investment yields returns. Examples of this are the movies “Evan Almighty” and “Super Troopers”. The former earned a gross of $100 million but cost $175 million to produce while the latter earned $18.5 million but cost $3 million to produce[5].

For this model, we have analyzed movies and found that production budget and genre are factors which are critical when determining whether the movie will break even or not in the opening week and resulting into high-profit margins[1]. Our motivation for this project is to build upon previous work related to movie revenue prediction and give investors more data to use and get more accurate results, hence improving the way in which investors can make a meaningful decision. Our model will give early predictions of movie profitability based on release date, production company, budget and genre.

For this project, we used data mining techniques, which are responsible for analyzing large quantities of data. In particular, we have used decision trees, a predictive model which is widely used. The two decision tree algorithms are based on Gini Index and Entropy and we will compare the two in order to determine the most efficient method. The decision tree algorithms discussed in this paper will extract useful and interesting information which can be used to develop decision support mechanisms for film producers. The first half of the report will discuss the related work and how the data was collected, cleaned up and processed.

## Related work

Large quantities of data regarding movies are generated and stored for analytical reasons and this shows the agency in the movie industry. The way in which success is defined is of paramount importance to the problem, but past works have focused primarily on gross box office revenue while some used the number of admissions. There are several related works involving the prediction of movie success based on reviews and box office. The basic assumption for using the two as success metrics is simple, a movie that sells well at the box office is considered a success. However, the two metrics ignore how much it costs to produce a movie. Google [2]; however, has created an application that predicts box office revenue

previous to opening weekend based on the search volume of the movie's trailer. The main difference between these related works and ours is that none of them predict movie success before the production begins. In fact, our analysis of data also found that budget is not directly related to profits. After a success metric was chosen, many studies categorized movies into two classes based on revenues either success or not and adopted binary classifications as their predictive task; we considered the prediction as a multi-class classification problem and tried to classify movies into several discrete categories. The most prominent include research described by Simonoff and Sparrow [3], using regression techniques to predict revenue from movies. More recent work has addressed similar research to the investigation reported in this paper, like, for example, research by Im [4] using a linear gradient descent algorithm to predict whether or not films will be profitable.

## Data

For the research presented here, we collected data from tmdb for 5000 movies that were released from 2005 to 2014 in the United States. We selected the data because it consists of both recent movies and old movies for which all the required data were available at the time of writing.

| company | release | genre | prod_budget | revenue |
|---|---|---|---|---|
| Other | 12 | Action | 3 | 5 |
| Walt Disney Pictures | 5 | Adventure | 3 | 5 |
| Columbia Pictures | 10 | Action | 3 | 4 |
| Other | 7 | Action | 3 | 5 |
| Walt Disney Pictures | 3 | Action | 3 | 2 |
| Columbia Pictures | 5 | Fantasy | 3 | 5 |
| Walt Disney Pictures | 11 | Animation | 3 | 3 |
| Other | 4 | Action | 3 | 5 |
| Warner Bros. | 7 | Adventure | 3 | 5 |
| Other | 3 | Action | 3 | 4 |
| Other | 6 | Adventure | 3 | 2 |
| Other | 10 | Adventure | 2 | 3 |
| Walt Disney Pictures | 6 | Adventure | 2 | 5 |
| Walt Disney Pictures | 7 | Action | 3 | 1 |
| Other | 6 | Action | 2 | 3 |
| Other | 5 | Adventure | 3 | 2 |
| Paramount Pictures | 4 | Science Fiction | 2 | 5 |

| Company | Release | Genre | Budget | Revenue |
|---|---|---|---|---|
| Other | 11 | Comedy | 0 | 0 |
| Columbia Pictures | 2 | Drama | 0 | 0 |
| Columbia Pictures | 1 | Action | 0 | 0 |
| Other | 4 | Romance | 0 | 0 |
| Other | 9 | Mystery | 0 | 0 |
| Columbia Pictures | 6 | Fantasy | 0 | 0 |
| Other | 2 | Action | 0 | 0 |
| Other | 1 | Comedy | 0 | 0 |
| DreamWorks SKG | 9 | Thriller | 0 | 0 |
| Village Roadshow Pictures | 3 | Action | 0 | 0 |
| Other | 9 | Drama | 0 | 0 |
| Other | 10 | Drama | 0 | 0 |
| Columbia Pictures | 12 | Drama | 0 | 0 |
| Warner Bros. | 8 | Comedy | 0 | 0 |
| Miramax Films | 12 | Comedy | 0 | 0 |
| Other | 3 | Drama | 0 | 0 |
| Columbia Pictures | 7 | Comedy | 0 | 0 |
| Other | 4 | Comedy | 0 | 0 |
| Twentieth Century Fox Film Corporation | 12 | Comedy | 0 | 0 |
| Other | 9 | Adventure | 0 | 0 |
| Other | 4 | Adventure | 0 | 0 |
| Universal Pictures | 6 | Comedy | 0 | 0 |
| Other | 8 | Comedy | 0 | 0 |
| Other | 5 | Comedy | 0 | 0 |

Figure 1 shows a fragment of the training data (left) and the test data (right). The original raw data compiled about each film included its production company, genre, title, multiple production companies, actors, budget and gross.

This research considered all the variables that in principle provided relevant information and were freely available. In order to stay true to our goal of only considering factors that affect the revenue before and after a movie's release, we considered only the production company, genre, title, budget, and revenue. Using Python to extract, parse, and clean up what we retrieved, we remove unnecessary data like multiple production companies and rows from each table that did not have a complete tuple of information. The dataset was divided into training and test data with six budget brackets (Increments of 125 million, up to > 500 million, and 0) and six revenue brackets (Increments of 250 million, up to > 1 billion, and 0). Splitting into training and testing sets was done by random assignment. When the data was ready we used 50% of the movies as a training set and the other 50% was used as the testing set. Data involving money will be corrected for inflation and put into brackets. Release dates will be categorized simply by month. Genres and Production Companies will be parsed, and the first one in each list will be used as the category for that movie.

## Our Approach

Classification is the task of assigning objects to one of several predefined categories. In classification, there is a given set of sample records called the training dataset with each record containing attributes. An attribute can be numerical or categorical. One of the categorical attributes is called the classification attribute and its values are called class labels[2]. The class labels indicate the class to which a record belongs. For this project, the expected revenue will be divided into six class labels whereby a movie classified in the category 5 is the highest profitable movie, the one in the 1 has the least non-zero profit, and the one in 0 has no profit. During classification, the model is created for each class and used to classify future records which are not present in the training dataset. The objective of our classification is to use the training data set to build a model of the class label such that it can be used to classify unknown class labels for new incoming data or new potential movies to be produced.

A decision tree classifier recursively builds a model by portioning the training data set so that most or all the records in a partition have the same class label. The tree is constructed and it is applied to each tuple in the database and this results into a classification for that tuple. The decision tree performs classification as follows:

***Tree-building:***

In this phase, the algorithm starts with the whole data set at the root node and the data set is then partitioned into subsets using a splitting criterion. This process is repeated recursively for each of the subsets until each subset is sufficiently small or when it has members belonging to the same class. The nodes represent the test on an attribute, the branches represent the value of the test performed in the node from which they branch off and the leaf nodes represent the class labels.

The first algorithm (CART, using Gini Index) for the building is as follows:

***BuildTree*** *(data set S)*
*if all records in S belong to the same class*
  ***return***
*for each attribute Ai*
  *evaluate splits on attribute Ai*
*use the best split found to partition S into S1 and S2*
***BuildTree*** *(S1)*
***BuildTree*** *(S2)*

The second algorithm (ID3, using Entropy) for the building is as follows:

***ID3*** *(data set S, attribute set A)*
*create Node*
*if all records in S belong to the same class*
  *label Node with class*
*else if A is empty*
  *label Node with majority class in S*
*else*
  *choose attribute to split by*
  *NewA = A - attribute*
  *partition S by values of attribute*
  *for each partition*
    *if partition is empty*
      *label Node with majority class in S*
    *else*
      *insert* ***ID3****(partition, NewA) to Node*
***return*** *Node*

The learning dataset comprises of different aspects of 50% of the movies. The attributes such as Production company, Production budget of the movie, Genre, and Revenue have been taken into consideration. If the revenue is poor, the movie may not be a hit and will be classified in a lower class. Furthermore, if the revenue is good, the movie may be successful based on the credentials such as budget, genre, release, company and will be classified in a higher class.

**Splitting procedure**

There are many measures that can be used to determine the best way to split the record. The splitting criterion is determined by picking the attribute with the best chance of separating the remaining sample of nodes partition into individual classes. This attribute becomes the decision attribute at that node. The tree uses different measures of impurity to select the split attribute in order to construct the decision tree. In this project, we consider Gini index and Entropy. We used the tree classifier which is constructed by repeatedly splitting subsets of the node into descendant subsets, beginning with node itself. To split the node into smaller and smaller subsets we have to select the splits in such a way that the descendent subsets are always purer than their parents. An impurity function is a function φ defined on the set of all k-tuples of numbers ;

*(p(c1),p(c2),...,p(ck)) satisfying p(ci)  0 ∀i ∈ {1,...,k} and k1=1 p(ci) = 1*

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. For each branch in a split, you calculate the percent branch represents and that will be used for weighting. The formula is shown below [8]:

*Let pij = relative frequency of class j in Si*

*ginisplit(S) = Σi [ gini(Si) * ni/n ]*

*Where:*

*gini(Si) = 1 – Σj [pij]2*

Entropy is a measure in information theory, which characterizes the impurity of an arbitrary collection of examples. It is easier to think of entropy as “the expected information needed to classify a tuple.” [9]

*Entropy(S) =∑- Pi log2Pi*

*Where:*

*Pi is the probability of S belonging to class i.* [8]

Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits. So if there is the more uniform probability distribution, the greater the entropy. Entropy is then used to calculate information gain. [9]

*InfoGain(Attribute) = Entropy(S) - Σ((Sj/S) * (Entropy(Sj)))*

*Where:*

*Sj/S is the probability that S belongs to class Sj,*

*Entropy(Sj) is the entropy of the subset of S that satisfies class Sj.*

A large InfoGain(Attribute) tells us that we have much to gain by branching out on Attribute. Using this, we can iterate through all possible attributes to split by and choose the one that gives the largest gain in information.

With several methods of predicting movie profitability, we chose decision trees because the data prediction needs less effort from users. For example, some of the research work uses a regression model, therefore to overcome the scale difference between parameters they would require some form of normalization or scaling before it can fit a regression model and interpret the coefficients[5]. Such variable transformations are not required with decision trees because the tree structure will remain the same with or without the transformation.

**Empirical evaluation**

The performance evaluation compares the decision tree based on Gini index and the one based on the entropy. We tested our decision tree by comparing the rules generated by the trees' branches to the movies in the test sets. In the first implementation, we go through the tree branches, matching attributes and class labels until hitting a leaf, which gives a prediction based on the majority count that it holds for the target attribute, in other words, it holds a list of counts for each revenue and will return the one with the highest count. In our second implementation, we go through the tree, forming a list of rules. We then match the labels in each movie with the labels in each rule, eventually arriving at the "revenue" label. We then update the count of correct classifications when the revenue bracket numbers match. As a sample to better visualize, we generated a distribution for each attribute and the intention is to see how each attribute affects the allocation of a movie in a specific class label (see Figure 2).

These numbers decrease with additional features, likely because of increased variance because of some overfitting on the training set. However, all of the algorithms had an accuracy over 50%. Gini index has an average accuracy of 63.79% while Entropy has an average accuracy of 54.3%.

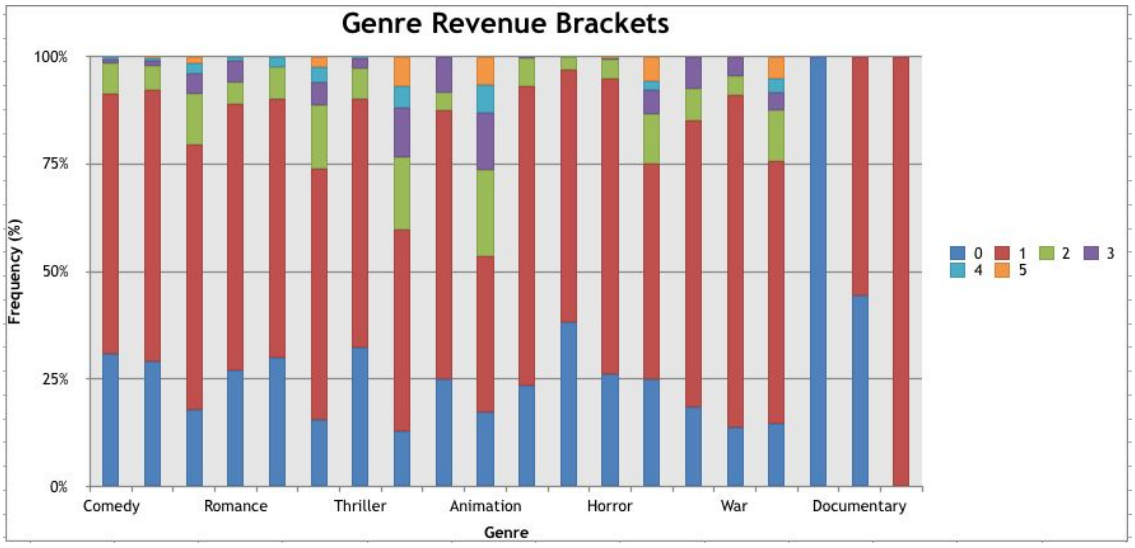


Figure 2 shows the distribution of the movies' revenue brackets according to genres, as an example.

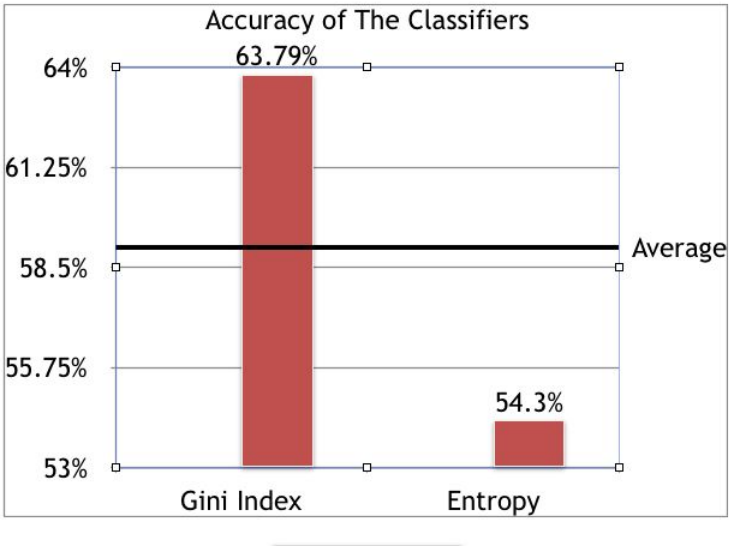


Figure 3 shows the difference in accuracy between using Gini index and Entropy/Information Gain.

There is room for improvement in order to get better accuracy. After building the decision tree, tree pruning can be implemented in order to get a smaller tree. Decision trees that are too large are susceptible to a phenomenon known as overfitting. As such, we highly suspect that implementing pruning would have allowed us to more accurately classify the test set of data, because a pruned tree tends to generalize more and overfit less.

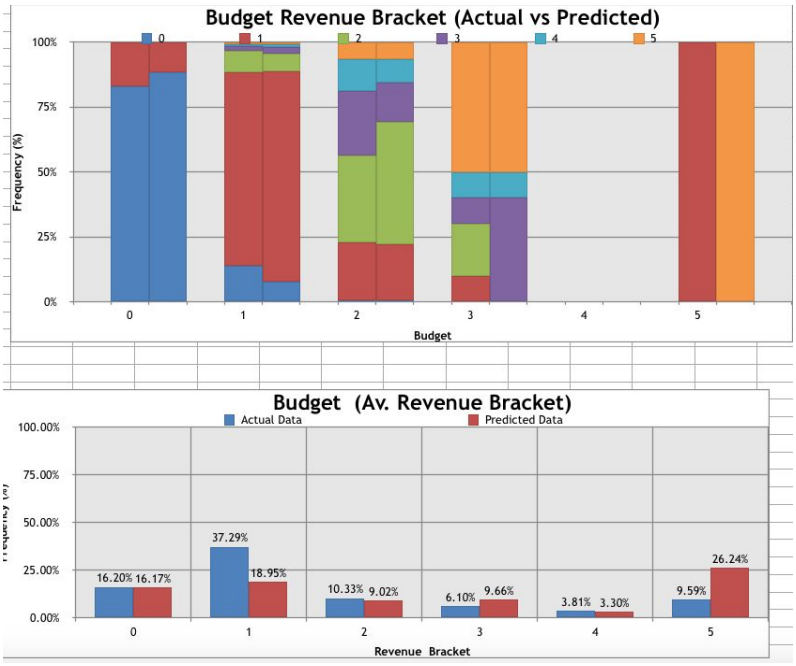


Figure 4 shows the differences between the actual data and the predictions from a decision tree.

We also noticed that training and test error rates of the model were large when the size of the tree is large, which was clearly a sign of overfitting. As a result, it performs poorly on both the training and the test sets. This is shown in Figure 4 and Figure 5. According to a researcher named Miniger, entropy is biased towards attributes with a large number of possible values. Miniger [1] compared entropy and $\chi 2$-statistic for growing the tree as well as for stop splitting. He concluded that $\chi 2$-corrected entropy's bias towards multi-valued attributes and this might be a possible reason why it did not perform as well as Gini index.

| | Frequency | | | | | | | Frequency (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Budget | 0 | 1 | 2 | 3 | 4 | 5 | Total | 0 | 1 | 2 | 3 | 4 | 5 |
| 0 | 323 | 66 | 1 | 0 | 0 | 0 | 390 | 82.82% | 16.92% | 0.26% | 0.00% | 0.00% | 0.00% |
| | 345 | 44 | 1 | 0 | 0 | 0 | 390 | 88.46% | 11.28% | 0.26% | 0.00% | 0.00% | 0.00% |
| 1 | 230 | 1251 | 137 | 30 | 13 | 14 | 1675 | 13.73% | 74.69% | 8.18% | 1.79% | 0.78% | 0.84% |
| | 132 | 1356 | 115 | 42 | 18 | 12 | 1675 | 7.88% | 80.96% | 6.87% | 2.51% | 1.07% | 0.72% |
| 2 | 1 | 33 | 50 | 37 | 18 | 10 | 149 | 0.67% | 22.15% | 33.56% | 24.83% | 12.08% | 6.71% |
| | 1 | 32 | 70 | 23 | 13 | 10 | 149 | 0.67% | 21.48% | 46.98% | 15.44% | 8.72% | 6.71% |
| 3 | 0 | 1 | 2 | 1 | 1 | 5 | 10 | 0.00% | 10.00% | 20.00% | 10.00% | 10.00% | 50.00% |
| | 0 | 0 | 0 | 4 | 1 | 5 | 10 | 0.00% | 0.00% | 0.00% | 40.00% | 10.00% | 50.00% |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.00% | 100.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% |
| | | | | | | | Actual Data | 16.20% | 37.29% | 10.33% | 6.10% | 3.81% | 9.59% |
| | | | | | | | Predicted Data | 16.17% | 18.95% | 9.02% | 9.66% | 3.30% | 26.24% |

Figure 5 shows the difference between training and test data.

Certainly, our choice of classifier and evaluation has limitations. For one, similar to past studies, the profit we calculated is based on estimated production budget and reported revenue. However, the true profit of a movie may be obscured by certain accounting practices. Gathering data that is accounted for the budget and revenue can be very challenging and that is beyond the scope of this paper. For many movies, theatre ticket sales are only one of the sources of income. Some movies may also rely heavily on the sale and rental of DVDs. A sample of several movies' actual revenue brackets shown with their predicted brackets can be seen on Figure 6.

| Company | Release | Genre | Budget | Revenue | Prediction |
|---|---|---|---|---|---|
| Paramount Pictures | 1 | Drama | 5 | 1 | 5 |
| Other | 3 | Family | 1 | 1 | 2 |
| Summit Entertainment | 9 | Thriller | 1 | 1 | 1 |
| New Line Cinema | 1 | Comedy | 1 | 1 | 1 |
| Other | 10 | Action | 0 | 1 | 0 |
| Other | 3 | Comedy | 0 | 0 | 0 |
| Other | 11 | Drama | 1 | 0 | 1 |
| Universal Pictures | 6 | Comedy | 1 | 1 | 1 |
| Touchstone Pictures | 6 | Adventure | 1 | 1 | 1 |
| Walt Disney Pictures | 12 | Romance | 1 | 2 | 1 |
| Other | 7 | Action | 2 | 4 | 2 |
| DreamWorks SKG | 12 | Drama | 1 | 1 | 1 |
| Other | 9 | Adventure | 0 | 0 | 0 |
| Other | 9 | Drama | 0 | 0 | 0 |
| Other | 9 | Science Fiction | 1 | 1 | 1 |
| Universal Pictures | 4 | Action | 2 | 5 | 2 |
| Other | 11 | Adventure | 2 | 5 | 3 |
| Twentieth Century Fox Film Corporation | 5 | Adventure | 2 | 2 | 4 |
| Universal Pictures | 8 | Drama | 1 | 2 | 1 |
| Other | 2 | Comedy | 1 | 0 | 1 |
| Lions Gate Films | 4 | Drama | 1 | 1 | 1 |
| Other | 8 | Thriller | 1 | 0 | 0 |
| Other | 6 | Adventure | 1 | 1 | 1 |
| Paramount Pictures | 10 | Documentary | 1 | 1 | 1 |
| Other | 3 | Crime | 1 | 1 | 1 |
| Other | 1 | Crime | 1 | 1 | 1 |
| Regency Enterprises | 2 | Comedy | 1 | 1 | 1 |
| Universal Pictures | 4 | Adventure | 1 | 1 | 1 |
| Universal Pictures | 4 | Horror | 1 | 1 | 1 |
| Other | 10 | Drama | 1 | 1 | 1 |

Figure 6 is a sample of movie information vs. our predicted revenue brackets.

**Conclusion and Future Work**

It is clear that predicting profit of a movie with a 100% accuracy can be difficult and with a large amount of data collected it becomes unclear which criteria are the best for-profit prediction. The project aims to predict movie's profitability before and after production in order to help investors and producer make a more informed decision where to invest in a movie or the effect of the budget on the retains from revenue. Our research aims to improve previous research by using a different type of classifier but based on previous related work, this might not be the case. For the rest of the report, we framed this problem to try and find the effective way to calculate the best splitting point using Gini index and Entropy. In general, we found that the decision tree based on the Gini index was a better classifier with an average accuracy of

63.79%, suitable to solve our problem. The lack of a pruning method proved to be a weakness in our implementations. For future work, we might consider using a pruning method in order to provide a more rigorous safeguard against high-variance or overfitting. This approach can be used as a support tool for prediction for upcoming movies.

**Reference**

[1] "Theatrical Market Statistics Report", Motion Picture Association of America, Inc., 2012. [Online]. Available: http://www.mpaa.org/wp- content/uploads/2014/03/2012-Theatrical-Market-Statistics-Report.pdf. [Access: 01-Apr-2016].

[2] K. Acuna, "Google Says It Can Predict Which Films Will Be Huge Box-Office Hits", Business Insider, 2016. [Online]. Available:

http://www.businessinsider.com/google-study-can- predict-success-of-movies-2013-6. [Accessed: 01-Apr- 2016].

[3] Simonoff, J. S., Sparrow, I. R., Predicting movie grosses: winners and

losers, blockbusters and sleepers. Chance, 13(3), 2000.

[4] Im, D., Nguyen, M. T., Predicting box-office success of movies in the

U.S. market, CS 229, Univ. Stanford, 2011.10

[5] Bozdogan, Y. The determinants of box office revenue: a case based study: thirty, low budget, highest ROI films vs. thirty, big budget, highest grossing Hollywood films. Master Thesis, University of Paris, 2013.

[6] Walls, W. D. Modeling Movie Success When Nobody Knows Anything: Conditional Stable Distribution Analysis Of Film Returns. Journal of Cultural Economics, 29, 3 (August 2005), pp 177–190.

[7] http://www.imdb.com/interfaces/

[8] Class Notes, 4710Unit10_Class.pdf: slide 15-16.

[9] Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Amsterdam: Elsevier/Morgan Kaufmann, 2012.