Liming Luo
Justin Murray
Youchen Ren

CS 780 Machine Learning
Project Report
May 23, 2014

Use of Neural Networks to Predict U.S. Movie Box Office Sales

I. Neural Network Background

An artificial neural network is modeled after that of a biological nervous system. It is an interconnected feed-forward network, which means a signal (or input) is received on one end of the model and then passed along through an indeterminate number of units (or layers) until it reaches the end (the output).

In an artificial neural network, there is an input layer, where the signal begins; one or more hidden layer(s); and finally an output layer, where the signal ends. Every layer starting from the input layer is fully connected to the next layer. Also, every connection from one layer to the next layer has a value associated with it, called a weight. These weights adapt during the training of the network, therefore allowing the network to learn as it receives more and more inputs. Every node within the neural network has an activation function assigned to it as well. It is this activation function that determines the output from the node to be carried to the next layer.

Backpropagation is a training algorithm that can be employed by a neural network. Back propagation networks learn by example; given some training data, the network modifies itself so that when training is completed, it will be able to give correct results for a test data point. The algorithm works in two steps: the Feed-Forward step and the Backpropagation step.

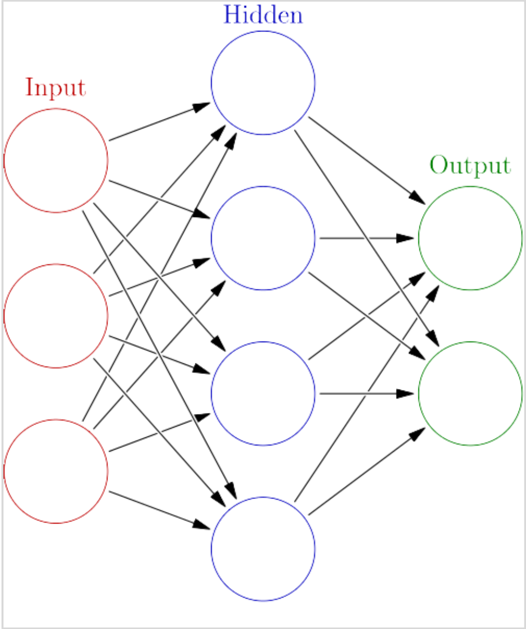


Figure 1 - Artificial neural network [1]

The Feed-Forward step will take an input and traverse it along the network and generate an output at the output layer. The Backpropagation step will take that output and calculate an error based on what the target value was on the original input. Based on this error, it will then modify the weights that are feeding into the node that produced the error. This is completed for all output nodes, and then the algorithm back propagates to the previous hidden layer of inputs and repeats the process of calculating an error and updating the corresponding weights. This process will continue until it reaches the input layer and then stops.

II. Test Model Construction

We constructed an artificial neural network to model the movie dataset that we compiled from OpusData.com. Our goal was to model the movie industry such that we could predict the profitability of a proposed movie given certain salient input features. The final network consisted of an input layer that takes in twenty-five features from the data set, a hidden layer with three hidden

Figure 2 - OpusData.com [2]

neurons, and an output layer with two output neurons. The neurons will be trained using the sigmoid function for activation and the error backpropagation algorithm on the target values provided by the data set.

The features from the dataset that were used as inputs can be seen in *Figure 3*:

nine creative types (index 0-8) valued as 1 if the movie was of that creative type and 0 otherwise; eleven genres (index 9-19) valued as 1 if the movie was of that genre and 0 otherwise; two actor profitabilities (index 20-21) see the *Further Development: Modeling Profitability* at the end of the report for explanation; one director profitability (index 22); the movie's budget (index 23); and whether the movie was a franchise or not (index 24) valued as 1 if the movie was a franchise and 0 otherwise.

0	Contemporary Fiction	Binary value
1	Dramatization	Binary value
2	Factual	Binary value
3	Fantasy	Binary value
4	Historical Fiction	Binary value
5	Kids Fiction	Binary value
6	Multiple Creative Types	Binary value
7	Science Fiction	Binary value
8	Super Hero	Binary value
9	Adventure	Binary value
10	Black Comedy	Binary value
11	Comedy	Binary value
12	Concert/Performance	Binary value
13	Documentary	Binary value
14	Drama	Binary value
15	Horror	Binary value
16	Musical	Binary value
17	Romantic Comedy	Binary value
18	Thriller/Suspense	Binary value
19	Western	Binary value
20	Actor 1 Profitability	Real number
21	Actor 2 Profitability	Real number
22	Director Profitability	Real number
23	Budget	Real number
24	Franchise	Binary value

Figure 3 - Model Inputs [3]

All profitabilities and budget values were normalized to between 0.0 and 1.0.

Three hidden neurons were employed for this project. During development of the neural network, tests on several different models were used to determine how many neurons the network should have in the hidden layer. However, after constructing models with 3, 4, 5, 6, and 7 hidden neurons and testing using 5-fold cross validation of the dataset on each of these models, we discovered that the lowest testing errors we could get from these models were very close to each other (*see Figure 4*). Since we were not receiving significant gains in accuracy as we increased the complexity of the model, we decided to use only three neurons in the hidden layer to reduce computational cost.

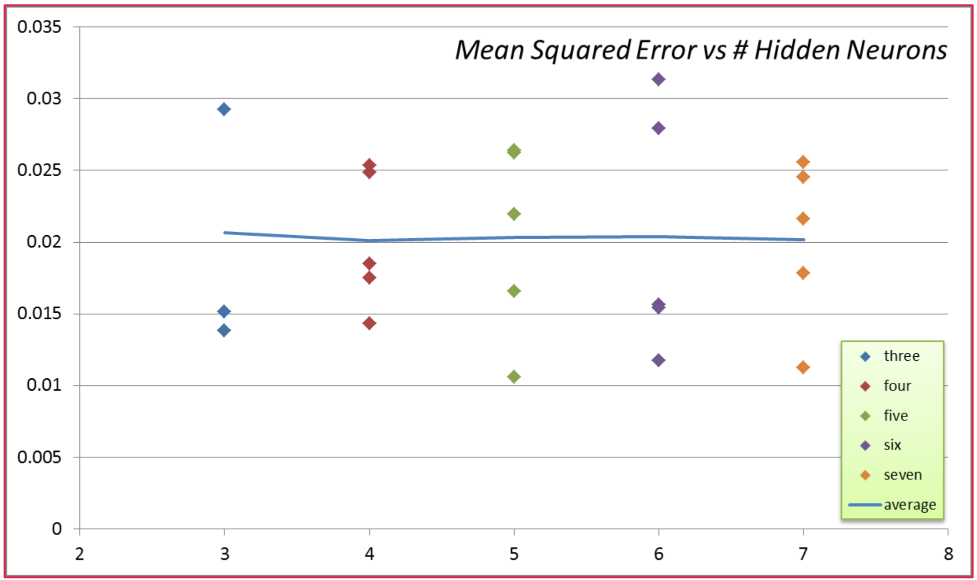


Figure 4 - MSE vs number of hidden neurons [3]

The two outputs (*see Figure 5*) are the target values for domestic and international box office revenue.

We constructed two separate models: the first with only one output neuron (domestic box office); the second with two outputs (domestic and international box office). We focused more on the latter model as that provided more interesting results.

0	Domestic Box Office	Real number
1	International Box Office	Real number

Figure 5 - Model Outputs [3]

We also ignored some features since they were not in the Opus database or we otherwise didn't have access to them, such as the release date, the quality of the script or competition from similar movies. These features will be discussed more in-depth in the further developments section.

III. Model Results

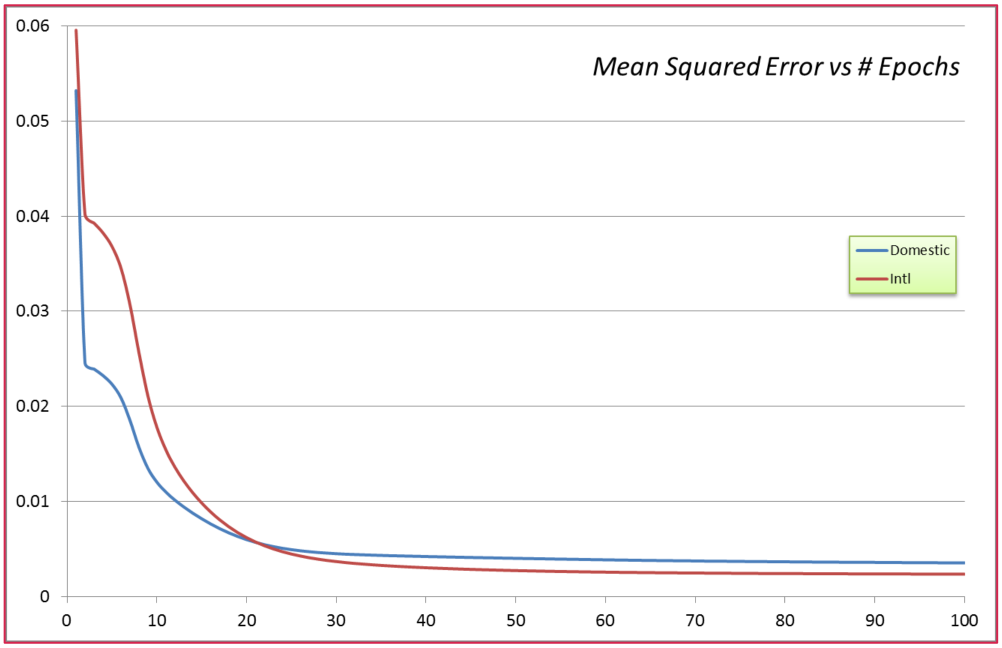


Figure 6 - MSE vs number of epochs for domestic and International box office [3]

To test the performance of the neural network, we used 301 data points and performed 5-fold cross validation. We used the delta of Mean Squared Error (MSE) on the training data as a stopping condition for our training of the network. Another stopping condition that was used was the maximum number of epochs allowed for each training, which was capped at 50,000 epochs. This caused all folds to be stopped at 50,000 epochs, even though the mean squared error was still declining at the end (*see Figure 6*). We had an average MSE across all folds of 0.0024 and 0.0017 for domestic and international box office, respectively (*see Figure 7*).

Training Cross Validation MSE

Fold	**Domestic**	**International**
1	0.002542311	0.001663041
2	0.002403202	0.001954428
3	0.002720745	0.001312682
4	0.001634885	0.001327619
5	0.002873499	0.001992556

Figure 7 - MSE for training data [3]

We then tested each fold with its corresponding testing data. The testing data during cross validation had an average MSE of 0.0088 for domestic box office, 0.0058 for international box office, and 87.0% accuracy (*see Figure 8*). Accuracy was calculated using the formula below:

$$Accuracy = \frac{number\ correct}{total\ number\ in\ sample}$$

An output was determined correct if it was within 0.1 of the target value. After analyzing the cross validation results, we decided on using the weights from fold 4 as they had the lowest combined MSE for both domestic and international box office.

Before we could start using those weights, we wanted to verify the ability of these weights to produce valid results. To verify this, we took three known movies and put them into the model and analyzed their results. We decided on two movies (*Avengers* and *World War Z*) that had high budgets and high box office totals, and one movie (*The Son of No One*) that had a small budget with poor box office performance (*see Figure 9*).

Testing Cross Validation MSE			
Fold	**Domestic**	**International**	**Accuracy**
1	0.0070688	0.0065064	90.0%
2	0.0046402	0.0039898	88.3%
3	0.0080779	0.0071539	81.7%
4	0.0214958	0.0071725	81.7%
5	0.0025648	0.0043622	93.3%

Figure 8 - MSE for testing data [3]

	Avengers		*World War Z*		*The Son of No One*	
Actors	Robert Downey Jr.	$1,151	Brad Pitt	$186	Channing Tatum	$70
	Chris Evans	$760	Mireille Enos	$350	Tracy Morgan	$14
Director	Joss Whedon	$1,289	Marc Forster	$350	Dito Montiel	-$14
Budget		$225		$190		$15
Creative Type		Super Hero		Science Fiction		Contemporary Fiction
Genre		Adventure		Action		Drama
Franchise		Yes		No		No

$ figure in $1,000,000

Figure 9 – Avengers, World War Z, and The Son of No One movie criteria [3]

Our model was able to predict the high budget movie's box office totals within reason: with only a 2.81% error for *World War Z*'s domestic box office total and a 13.53% error for international box office (*see Figure 10*). Unfortunately, our model was not able to predict if a movie bombs at the box office, in the case of *The Son of No One*. One possible explanation could be the disproportionate effect a director's profitability has on the movie, which will be discussed later.

	Domestic			**International**		
	Computed	**Actual**	**% Error**	**Computed**	**Actual**	**% Error**
Avengers	$570,581,206	$623,279,547	**8.46%**	$748,161,854	$891,000,000	**16.03%**
World War Z	$196,675,297	$202,359,711	**2.81%**	$291,560,290	$337,200,000	**13.53%**
The Son of No One	$16,634,508	$30,680	**54119.39%**	$13,607,587	$1,117,898	**1117.25%**

Figure 10 – Results from the model using the Avengers, World War Z, and The Son of No One movie criteria [3]

With our results from our five-fold cross validation producing low mean squared errors, very nice looking trends of MSE vs epochs, and respectable results on our analysis of *Avengers*, *World War Z*, and *The Son of No One*, we then created a movie and tested the various inputs to see their effect on a movie's profit or loss.

IV. Movie Input Analysis

For our movie, we randomly selected two actors (*Arnold Schwarzenegger* and *Jennifer Lawrence*) and a director (*Christopher Nolan*). We then set the movie's budget to $50 million, creative type to contemporary fiction, and genre to romantic comedy. Our model predicted that the movie would gross over $1 billion worldwide and have a profit/loss ratio (PLR) of 20.718 (*see Figure 11*). PLR was calculated using the formula below:

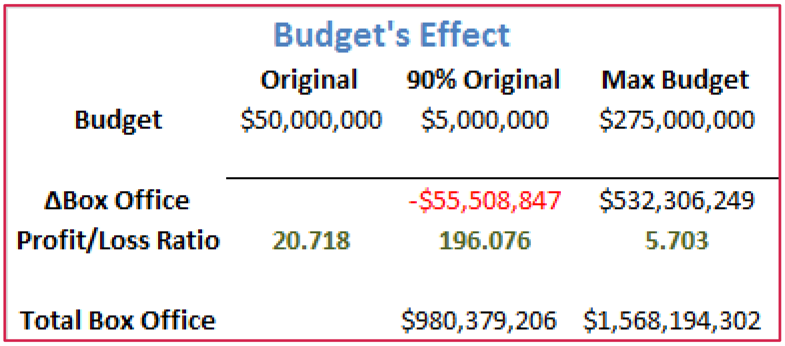
Arnold Schwarzenegger	$25,931,701
Jennifer Lawrence	$512,832,045
Christopher Nolan	$804,343,943
Budget	$50,000,000
Creative Type	Contemporary Fiction
Genre	Romantic Comedy
Domestic	$125,042,318
International	$910,845,735
Total Box Office	**$1,035,888,053**
Profit/Loss Ratio	**20.718**

Figure 11 – Created movie criteria [3]

$$\text{Profit/Loss Ratio} = \frac{\text{total box office}}{\text{budget}}$$

This ratio will be an easy indicator for us to determine if a change in the movie parameters will be beneficial to the investor or not. One of the first inputs we thought would have a major impact on the box office sales of a movie was the movie's budget (*see Figure 12*).

Our original movie had a budget of $50 million. Our model shows us that a reduction of our budget by 90% will result in $55 million less than our original movie's box office sales, but our PLR goes up almost 4 times. For every dollar we invest into the movie, we would get $196 back, whereas the original budget would only net a return of $20.72 on the dollar. Also, if we were to max out the budget to $275 million, our box office sales increase over $530 million but the PLR drops to 5.703.



Budget's Effect

	Original	**90% Original**	**Max Budget**
Budget	$50,000,000	$5,000,000	$275,000,000
ΔBox Office		-$55,508,847	$532,306,249
Profit/Loss Ratio	20.718	196.076	5.703
Total Box Office		$980,379,206	$1,568,194,302

Figure 12 – Effect of budget has on a movie's profit/loss ratio [3]

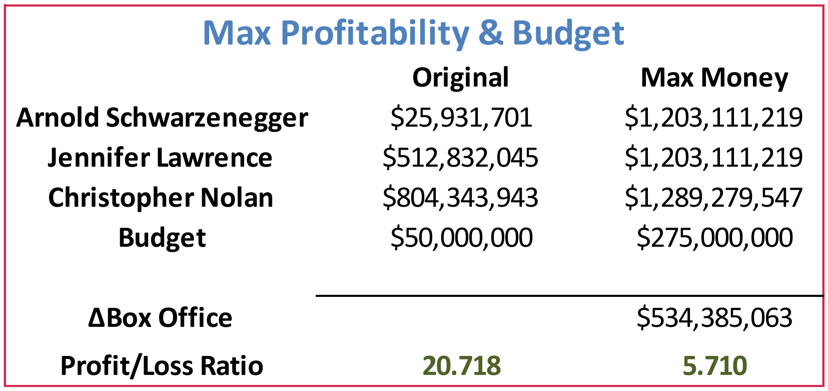
Next, we wanted to test the importance of the actor's and director's profitability (*see Figure 13*). These results seem to go against conventional wisdom. Our analysis on our model has shown us that a director's profitability has the single largest impact on a movie's box office sales. If the movie studio were to replace highly profitable actors with actors that have virtually zero

Profitability

	Original	**Unknown Actors**	**Unknown People**	**Max Profitability**
Arnold Schwarzenegger	$25,931,701	$1	$1	$1,203,111,219
Jennifer Lawrence	$512,832,045	$1	$1	$1,203,111,219
Christopher Nolan	$804,343,943		$1	$1,289,279,547
ΔBox Office		-$24,519,434	-$982,169,449	$533,644,733
Profit/Loss Ratio	**20.718**	**20.227**	**1.074**	**31.391**

Figure 13 – Effect of profitability on a movie's profit/loss ratio [3]

profitability, the effect is minimal as the PLR only drops by 0.491. But if the movie studio replaces a highly profitable director with someone with a tiny profitability, the PLR drops to an astonishing 1.074; just above the break-even mark. In contrast, if the movie studio replaces the actors and director with the most profitable people on the market, the PLR will jump over 10 points.

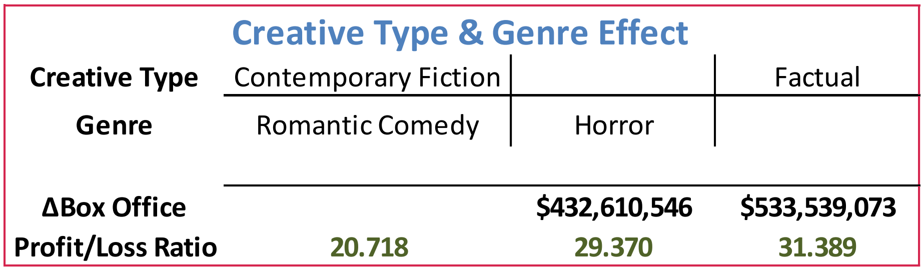


Max Profitability & Budget

	Original	Max Money
Arnold Schwarzenegger	$25,931,701	$1,203,111,219
Jennifer Lawrence	$512,832,045	$1,203,111,219
Christopher Nolan	$804,343,943	$1,289,279,547
Budget	$50,000,000	$275,000,000
ΔBox Office		$534,385,063
Profit/Loss Ratio	20.718	5.710

Figure 14 – Effect of profitability and budget on a movie's profit/loss ratio [3]

After analyzing the importance of a movie's budget, actor's profitability, and director's profitability, we then analyzed the importance these have on a movie combined (*see Figure 14*). When we supplied a movie with the most profitable actors and director, along with a max budget, the PLR actually drops to ¼ of what it once was, even though the movie will increase its box office sales by over $530 million.



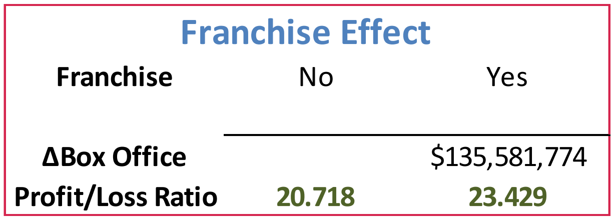
Creative Type & Genre Effect

Creative Type	Contemporary Fiction		Factual
Genre	Romantic Comedy	Horror	
ΔBox Office		$432,610,546	$533,539,073
Profit/Loss Ratio	20.718	29.370	31.389

Figure 15 – Effect of creative type and genre on a movie's profit/loss ratio [3]

The final three inputs that were tested do not deal with money directly. Creative type and genre were analyzed next. When a movie's genre is changed from romantic comedy to horror, the movie was shown to perform better, increasing its PLR from 20.718 to 29.370. A change of creative type from contemporary fiction to factual has an even greater effect on PLR, raising it to 31.389 (*see Figure 15*). The importance of each creative type and genre individually should be investigated in the future.

Conventional wisdom tells us that if movie is a part of a franchise it should perform better than a non-franchised movie for at least two reasons. First, the movie studio would not produce a second movie if the first movie did not perform reasonably well in the first place. Second, people will feel inclined to see a sequel if they have seen the previous movie because they are invested into the franchise. Our model proves these assumptions to be correct. A franchised movie performs slightly better than a non-franchised movie (*see Figure 16*).



Franchise Effect

Franchise	No	Yes
ΔBox Office		$135,581,774
Profit/Loss Ratio	20.718	23.429

Figure 16 – Effect of franchise on a movie's profit/loss ratio [3]

V. Further Development

1. *Dataset*

Our dataset included only movies in the past four years with a budget of at least $1 million and had domestic box office sales of $10,000 or more. For a more accurate model, this dataset should be expanded to include smaller budget movies in a longer timeframe. In particular, with our current dataset we had some problems with small sample sizes for some of the less popular genres (ex: horror) which skewed our results to make these genres seem more or less profitable than they really were; a larger dataset would solve these problems. Further, including lower-budget movies would give us a more flexible model that would allow us to make predictions about movies developed by smaller studios.

2. *Modeling profitability*

Our current profitability is calculated as follows:

$$\left(\frac{1}{N}\right)\sum_{i=1}^{N}(total\ box\ office\ revenue_i - budget_i)$$

where N is the number of movies the person appears in the credits

As actor and director profitability is such an integral part of our model, it seems only reasonable that further development focus on a better way to calculate this profitability. Our equation was very simple and only considered movies within the date and budget parameters of our dataset for any individual actor or director, but a more sophisticated model for profitability might take into account the person's entire film history, the quality of their work (perhaps judged by the number/prestige of the awards they've received), the size of their fan base, and so on. Indeed, modeling actor or director profitability may very well be suitable for a machine learning project on its own.

3. *Feature selection*

Several important features were omitted or simplified due to time constraints and availability of data. Further development would add these features to our model, such as:

a. **Release date**: it is already widely known in the movie industry that the timing of a movie's release has a large effect on its box office revenue, leading to certain seasonal release patterns (for example: action blockbusters during the summer; family films during the holiday season; etc.) which we did not have the data to examine. Including this feature would also allow us to then compare the profitability of a movie across different release times, particularly coupled with genre information.

b. **Competition from similar releases**: if two movies of the same "type" (similar genre and creative type) are released very close to each other, a customer will likely pick one or the other but probably not both, thus affecting the revenue for both movies.

c. **Budget**: this feature was included in our model inputs, but it was treated as one lump sum, and could be further broken down into production costs and advertising costs.

d. **Runtime**: does the length of a film have an effect its profitability?

e. **Release pattern**: different release patterns (wide, limited, delayed, etc.) mean different levels of access to a movie.

f. **Script quality**: this would be very difficult to gauge (perhaps it could be measured by writer profitability?) but the quality of a movie script certainly contributes to its success.

4. *Model Structure*
   a. **Output layer**: Our final model had only two neurons for the output layer—domestic box office revenue and international box office revenue. However, some other useful outputs to consider may be gross revenue (including DVD sales, merchandising, etc.) and critical reception.

   b. **Hidden layer**: As mentioned previously, 3 hidden neurons may not be the optimal choice for our neural network. Further development would experiment with the number of hidden neurons to try and increase the accuracy of the model.

References:

1. "Artificial neural network - Wikipedia, the free encyclopedia." 2004. 22 May. 2014 <http://en.wikipedia.org/wiki/Artificial_neural_network>

2. "OpusData." 2002. 22 May. 2014 <http://www.opusdata.com/>

3. "final_results.xlsx" 23 May. 2014 <http://drive.google.com/open?id=0Bwufg68K4B_rWWh6WlFDNldHam8>