Team Members:
Yingda Hu (yhx640), Brian Zhan (bjz002)
James He (jzh642), Kevin Cheng (klc954)

Machine Learning Final Report- EECS 349 Spring 2016
Instructor: Doug Downey

Project Goal:

The goal of our project is to predict the opening week box office revenue of a movie based upon features found from its YouTube trailer. This is an important task for the multi-billion dollar movie industry where producers often compete for opening weekend box office sales. Despite the literature on this topic, it is not uncommon for experts to incorrectly predict opening box office sales. Past studies in this field like "Pre-production Forecasting of Movie Revenues with a Dynamic Artificial Neural Network" state that word of mouth may be used to improve forecasts, but they subsequently fail to measure this. Furthermore, other studies frequently discretize opening box office to excessively large ranges: "Predicting box-office success of motion picture with neural networks", for instance, attempts to predict box office revenue in the ranges of 40-65 million, 65-100 million, and 100-150 million--such a large range makes the results difficult to use for practical purposes. In this study, we hope to add YouTube data as a measure of "word of mouth" to the overall forecasting.

Overview of our Dataset:

Our data set consists of data from about 300 different movie trailers between the years 2010 and 2016, independently gathered from one Youtube channel Movieclips Trailers to reduce the amount of viewer bias that occurs between popular and unpopular channels. We kept the movies recent (after 2010) because as the Internet grows, more and more people are using Youtube for movie trailers, which might potentially skew the number of viewers towards newer movies.

The data set has a total of 9 attributes. These attributes are: the number of views (adjusted for number of days ahead of movie release the trailer is released), times shared, like:dislike ratio, movie genre, number of theaters showing the movie, and the opening box office revenue.

We wrote a script to assist in the collection of this data; it takes in a movie trailer url and returns the number of trailer views prior to movie release date, ratio of likes to dislikes, and the number of times the trailer was shared. Then, we manually filled in the genre of each movie, the number of theaters the movie was released in, and the opening week box office revenue with the assistance of the website www.boxofficemojo.com

After analyzing our dataset, we noticed that our data contains fewer trailers that had above 60 million dollars in opening box office revenue. This was mostly because there were not many movies with a high enough opening box office revenue to start out with.

Initial Approach:

We first ran our data set using ZeroR to obtain a baseline accuracy. ZeroR reported an accuracy of 57.3123%.

Then, we experimented with training on both discretized and un-discretized data with 10-fold CV. Our results are displayed in the table below. (note: at this stage, there were 300 movies in data)

Discretized		**Non-Discretized**		
Algorithm	**Accuracy;** **Mean Absolute Error** **($10 million)**	**Algorithm**	**Correlation Coefficient**	**Mean Absolute Error** **($million movies)**
J48 Pruned Decision Tree	**61.194%**; 0.0695	Decision Tree: REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1	**0.6631**	$9,810,427
MLP Neural Network (LR=0.3; M=0.2)	**62.866%**; 0.0614	MLP Neural Net (L=0.3 M=0.2)	**0.5238**	$15,504,713
Bayesian Network (Simple Estimator -A 0.5; Search Algorithm: K2 -P)	**64.9254%**; 0.0619	LinearRegression -S 0 -R 1.0E-8	**0.5815**	$14,557,610
Nearest Neighbors (IB-1)	**55.9701%**; 0.0545	Nearest Neighbor: IBk -K 1 -W 0 -A	**0.4621**	$14,661,161

On the discretized set, it is unsurprising that bayesian and neural networks produced the best results: most modern literature either use econometric models or these networks to forecast movie box office sales. This is because they have the ability to learn regression curves to forecast accurately. On the non-discretized set, neural/bayesian networks performed surprisingly poorly--this is likely because they require more data to forecast precise box office revenues.

From the above results, we determined that the discretized data produced better results, so we focused on trying to optimize the accuracy for discretized data.

Improving The Accuracy With Data Processing and Feature Analysis:

We first verified that the fact that all our trailers came from one YouTube channel would not introduce new bias. To do this, we compared the correlation coefficient between trailer data from the most popular trailer on YouTube to the correlation coefficient on trailers from MovieClips Trailer, the channel that we used. We found that MovieClips Trailers gave a slightly higher correlation coefficient, so our method is viable. However, we noticed that some trailers, for example “Captain America Civil War” had much higher view counts on the main channel it was released on, in this case “Marvel Entertainment”. To account for this form of bias, we made sure that our chosen trailer from the MovieClips channel was among the top 3 trailers for that given movie among all the other channels on Youtube. This would ensure that the number of views, shares, and likes would not be disproportionately lower due to the higher view count on the same trailer but on a different channel.

We then tried to measure the importance of our features, by removing one attribute at a time and seeing how it affects the accuracy. Through this, we determined that the number of days since trailer release and number of dislikes were unimportant features and removing them actually improved our prediction accuracy.

Finally, we scaled the number of views per trailer by dividing the number of views by log(# of days since trailer release) to account for the fact that trailers that have been released for a longer period of time would result in more total views.

Results:

Testing on Different Algorithms with 10-Fold CV (w/ 450 movies in data set):
The following table showcases the validation accuracies from the different machine learning algorithms utilizing our new processed data set. The accuracy outputted using the ZeroR method is our baseline accuracy, as no rigorous or in-depth machine learning techniques are used in the ZeroR method. The use of multilayer perceptrons through neural networks outputted the highest accuracy through 10-fold cross validation. This was expected based on other studies we read in the field of forecasting box office revenue: neural networks are able to learn “curves” of much less smooth data. We were surprised to see that nearest neighbors performed the worst, but this is likely explained by the existence of a lot of noise in our dataset, which has a particularly detrimental effect to the nearest-neighbor algorithm.

Algorithm	**10-fold CV Accuracy**
ZeroR	64.9254%
IB-k Nearest Neighbor	69.194%
MLP Neural Net (L=0.3 M=0.2; Epoches=450)	78.866%
Bayes Nets	72.727%
J48 Decision Tree	75.866 %

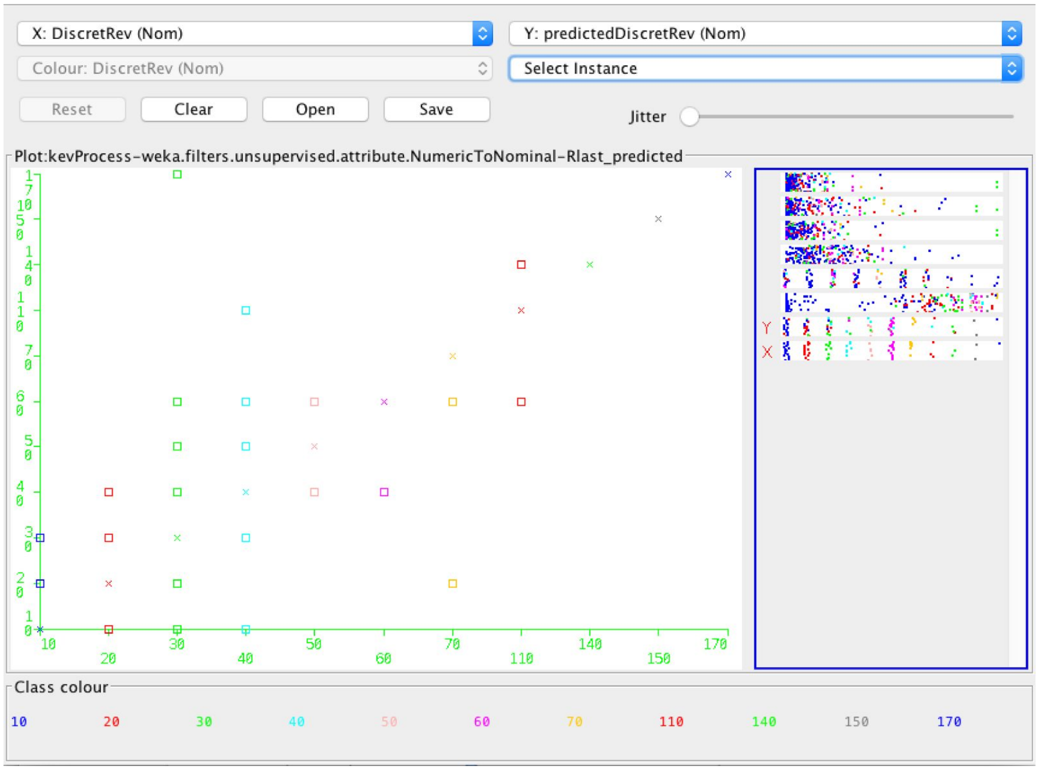
Testing on “Box Office Flops”:
We also tested our fully trained neural network on 10 “box office flops” (whose failure was a surprise): these included Doctor Dolittle, The Fall of the Roman Empire, Hello, Dolly!, Honky Tonk Freeway, Krull, The Last Castle, Nothing but Trouble, Peter Pan, The Scarlet Letter, and Soldier. The classification accuracy (into output ranges differing by $10 million) was 60%. While not a large data set, this shows that our model has learned with acceptable accuracy to predict some of the “unpredicted”.

Team Members:
Yingda Hu (yhx640), Brian Zhan (bjz002)
James He (jzh642), Kevin Cheng (klc954)

Analysis of Results:

Analysis of MLP, Most Accurate Model:

The following is a graph of our weka output of the multi-layer perceptron neural net output.



X: DiscretRev (Nom)

Y: predictedDiscretRev (Nom)

Colour: DiscretRev (Nom)

Plot:kevProcess-weka.filters.unsupervised.attribute.NumericToNominal-Rlast_predicted

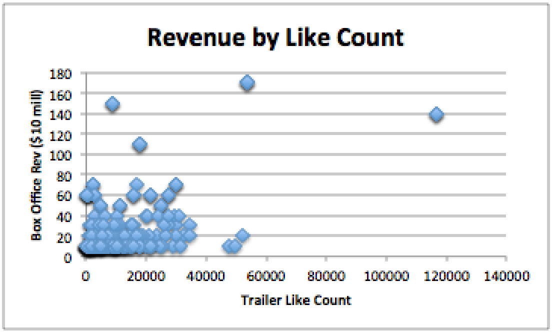
This graph displays the predicted discretized revenues vs the actual discretized revenues when putting the dataset through a MLP neural net. The x axis is the actual revenues and the y axis is the predicted revenues. Boxes show incorrect opening box office revenue predictions, while x's mark correct predictions. The abundance of mistakes in the $20-$30 and $30-$40 million range is partly due to the lack of movies with this opening box office revenue in our model (caused by the lack of movies in general with opening revenues above $30 million). While >$40 million range did not have as many mistakes, this is mostly because of less movies in our data set in this range.

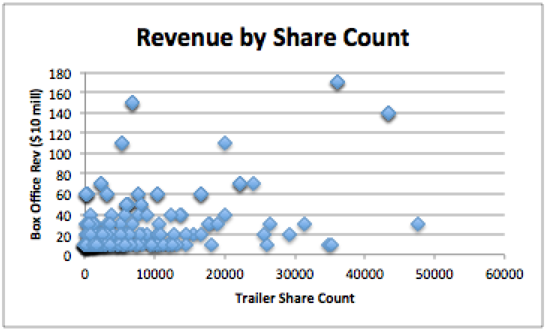
Neural Network Parameters: A learning rate of 0.3 and momentum of 0.2, with 500 epochs proved to be optimal: increasing or decreasing learning rate/momentum (w/ adjustments to epoches) both increased error, showing these configurations were best for learning features at appropriate pace whilst overcoming local optimums without overfitting. Optimal number of epochs was determined manually: Increasing number of epochs to 500 decreased the CV accuracy to 77.47%; decreasing number of epochs to 400 led to minor drop in accuracy to CV 78.101%; this shows that 450 epochs was approximately optimal to learn the features whilst reducing overfitting.

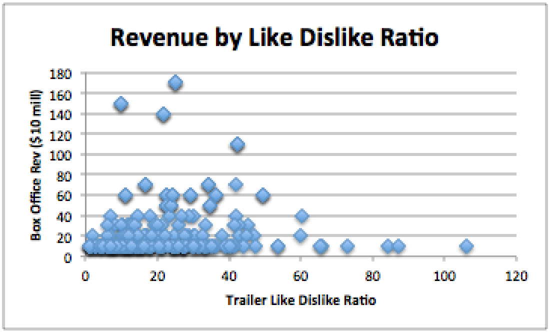
Most Important Features:

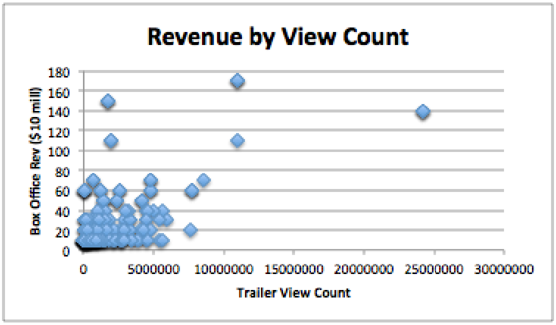
Team Members:
Yingda Hu (yhx640), Brian Zhan (bjz002)
James He (jzh642), Kevin Cheng (klc954)

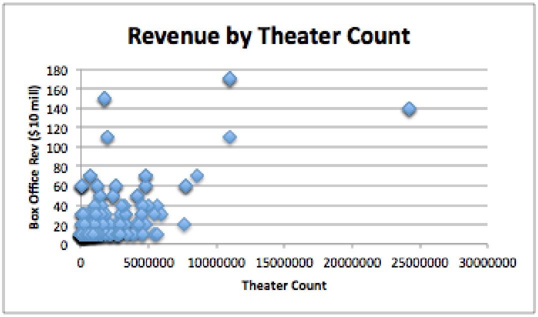
The neural network generated 23 sigmoid nodes. Overall, the weights ranked the following in importance: view count, share count, theater count, like count, like:dislike ratio, followed by weights for the movie genre, with view count and share count receiving significantly more weight than other attributes. This is not surprising, given that these two attributes showed stronger correlation to opening box office revenue, as shown in the below charts of our raw data.


Revenue by Like Count


Revenue by Share Count


Revenue by Like Dislike Ratio


Revenue by View Count


Revenue by Theater Count

Future improvements:

The movie industry is impacted by many different factors that can be difficult to capture and represent in our dataset. Moreover, the complexities of human nature and the relationship between different factors in daily life can also impact the success of a movie. To improve our dataset, we can run sentiment analysis on the comments section of a particular movie trailer (through the use of AI algorithms), which could potentially allow us to gauge the initial reactions to a specific movie. Furthermore, we can also create an evaluation function for different producers or actors that appear in a movie, giving a higher evaluation to more prominent actors/actresses or producers, as popular producers tend to have higher box office ratings (i.e. George Lucas).

Who did what:
Kevin Cheng - Worked on collecting data, running the data through Weka, writing the the report, creating the website
Yingda Hu- Wrote script to collect data from YouTube, assisted in manual collection of data and processing, analyzed usefulness of individual features, worked on report
James He - Worked on collecting movie data from Youtube, writing all the reports (proposal, progress report, etc.), and analyzing the data
Brian Zhan - Incorporated web scraping to collect data the YouTube API doesn't; assisted in manual collection of data and processing; optimized accuracy of machine learning algorithms; helped analyze neural network results

Works Cited:

Ghiassi, M., David Lio, and Brian Moon. "Pre-production Forecasting of Movie Revenues with a Dynamic Artificial Neural Network." *Expert Systems with Applications* 42.6 (2015): 3176-193. Web.

Sharda, R., and D. Delen. "Predicting Box-office Success of Motion Pictures with Neural Networks." *Expert Systems with Applications* 30.2 (2006): 243-54. Web.