Using Sentiment and Social Network Analyses to Predict Opening-Movie Box-Office Success

by

Lyric Doshi

B.S., Massachusetts Institute of Technology (2008)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the Massachusetts Institute of Technology

February 2010

Author .....
Department of Electrical Engineering and Computer Science
February 2, 2010

Certified by .....
Peter Gloor, Research Scientist
MIT Thesis Supervisor

Accepted by .....
Christopher J. Terman
Chairman, Department Committee on Graduate Theses

Using Sentiment and Social Network Analyses to Predict Opening-Movie Box-Office Success

by

Lyric Doshi

Submitted to the Department of Electrical Engineering and Computer Science
on February 2, 2010, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

In this thesis, we explore notions of collective intelligence in the form of web metrics, social network analysis and sentiment analysis to predict the box-office income of movies. Successful prediction techniques would be advantageous for those in the movie industry to gauge their likely return and adjust pre- and post-release marketing efforts. Additionally, the approaches in this thesis may also be applied to other markets for prediction as well.

We explore several modeling approaches to predict performance on the Hollywood Stock Exchange (HSX) prediction market as well as overall gross income. Some models use only a single movie's data to predict its future success, while other models build from the data of all the movies together. The most successful model presented in this thesis improves on HSX and provides high correlations/low predictive error on both HSX delist prices as well as the final gross income of the movies. We also provide insights for future work to build on this thesis to potentially uncover movies that perform exceptionally poorly or exceptionally well.

MIT Thesis Supervisor: Peter Gloor
Title: Research Scientist

Acknowledgments

I would like to thank Peter Gloor of MIT's Center for Collective Intelligence for being a great adviser and mentor to me on this project, providing great input and insight as my work progressed. Peter was very encouraging throughout, sharing and adding to my excitement for all my results, both the small and the meaningful. Peter's understanding and support made my thesis experience very educational and enjoyable at the same time.

I also want to thank Thomas Malone of MIT's Center for Collective Intelligence for taking time out of his very busy schedule first to help bring me into this project and then to provide key insights and encouragement as my work progressed.

My colleagues Jonas Krauss, Stefan Nann, and Hauke Führes from the University of Köln were a pleasure to work with. I have to thank them for contributing key components such as including some of the web crawlers, the sentiment analysis scripts, and the social network analysis modules. I also want to specifically thank Jonas and Stefan for repeatedly tweaking some of the sentiment scripts for me even on short notice. Finally, I must thank them for the amazing hospitality the three of them showed me when I took an impromptu week-long trip to Köln, Germany, while working on this project.

I need to thank my sister Finale Doshi, currently a Ph. D. student at MIT, for spending many hours explaining and sometimes re-explaining various statistics and modeling concepts to me as related to my thesis and attempts to build good movie predictors.

Finally, thank you to Simone Agha for reading over my thesis and finding my late-night typing mistakes and inventive grammar constructions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The purpose of this research is to explore the effectiveness of collective intelligence, social network analysis and sentiment analysis in predicting trends by mining publicly available online data sources. We aim to garner the opinions of the movie-watching public on the Web to make meaningful predictions by putting together the little pieces of the big picture each person can provide. More specifically, this research focuses on predicting the success of new movies over their first four weeks in the box office after opening as a medium to explore these ideas.

This thesis will describe previous work with making predictions and why we believe using concepts of collective intelligence may make an impact. We will describe our exploration of several approaches for making predictions as we build up to our most successful attempt described in Section 5.2.2. Finally, we offer suggestions for further research.

1.1 Motivation

The driving motivation behind the movie success prediction is two-fold. First, having a strong estimator of a movie's anticipated performance around release time can influence roll-out and marketing decisions to increase movie revenue. The movie prediction problem also provides a relatively controlled environment to explore and test prediction algorithms for the general case. It barely needs to be said that high-

confidence predictions would be extremely helpful and lucrative in decision-making across many disciplines, including predicting election and stock market outcomes. In both cases, we hope to explore and demonstrate the power of collective intelligence in making such predictions.

1.2 Goal

To narrow down the meaning of predicting box office success, we pursue two different goals in this thesis.

First, we try to anticipate the day-to-day fluctuations in the revenue predictions of the Hollywood Stock Exchange prediction market. That is, we predict what other people predict each movie will gross in revenue. Success here would provide an oracle-like tool to develop investment strategies. Given that some applications such as stock market prediction have this form of behavior – daily fluctuations with no clear endpoint in sight – there are clear benefits for accurate prediction here.

Second, we make longer-term predictions of the movies' actual final gross revenue in the box office. This goal is more in line with our specific goal of determining how well movies will perform, since it is the end result and not peoples opinions that matter in the end. The existence of a definite ending point time-wise to the prediction period along with a defined value changes the nature of the problem a little, but we hope it makes it more tractable as well. We hypothesize that this second approach can be considered a special case of the first problem with the price of a certain day considered as the "end point."

1.3 Contributions

In this thesis, we make the following contributions:

1. Suggest a new, potentially powerful metric in betweenness centrality for networks built on websites describing a movie instead of the typical social networks for which this metric is commonly used

2. Describe and explore several modeling approaches to predict HSX and box-office gross income success

3. Show that HSX can be used to predict not only the four-week income of a movie as it is designed, but also the final gross income of the movie

4. Demonstrate a method to improve on the already high-quality predictions from the HSX market

5. Provide evidence suggesting that the number of people with strong opinions correlates well with the success of a movie, regardless of whether those opinion are strongly positive or strongly negative

6. Give suggestions to build on and extend this work, including a method to potentially predict outliers that perform very well or very poorly

1.4 Outline

In Chapter 2, we give a review of some previous work in this area. Chapter 3 describes the resources we used to garner data for the predictions. Next, we discuss the models built using only a single movie's data to predict its future success in Chapter 4. Chapter 5 describes the models built by combining all the movies data together into a single, more general model for movie success rather than a model per movie. Our most successful efforts can be found here. Chapter 6 gives some thoughts for future work to build on this thesis and Chapter 7 concludes the thesis.

Chapter 2

Background and Related Work

2.1 Collective Intelligence

Collective intelligence has been defined by Thomas Malone of MIT's Center for Collective Intelligence to be "groups of individuals doing things collectively that seem intelligent" [9]. He goes on to explain that "Collective Intelligence relies upon the individual knowledge, creativity, and identity of its constituent parts, and emerges from a synergy between them. In its highest forms, participating in collective intelligence can actually help people self-actualize while solving collective problems." With these definitions, Malone seeks to distinguish collective intelligence from individuals acting together to generate false consensus, cults, hive minds, or Groupthink. Collective intelligence is about making more informed decisions and conclusions based on the contributions of specific knowledge or expertise of many, not acting as an impulsive or emotional mob.

Bonabeau explains in the MIT Sloan Management Review that collective intelligence can allow for better outreach, aggregation, and organization of ideas [3]. However, a proper collective intelligence system must also address issues including the loss of control from an administrator and the policing of contributions to ensure they are valid. The system must engage participants to contribute and balance the question of having diversity of opinion of many against the opinion of those with expertise in the matter at hand. The quality and popularity of Wikipedia has been a testament

to the power of collective intelligence. Prediction markets, too, have been excellent in capturing the collective input of participants with their own external and insider information to make meaningful conclusions and predictions.

2.2 Prediction Markets

Prediction markets tie payoffs with the outcomes of future events. The goal is for the payoffs to be linked with the likelihood of the outcomes based on the knowledge and insights of the participants. As Wolfers and Zitzewitz describe in [18], there are three main types of prediction markets. The first is a "winner-take-all" market where a contract pays a fixed amount only if some event occurs. The price for the contract varies with how likely people think the event will happen. Thus, it reveals the market expectation that the event will occur. The second type is an index market. Here, the contract pays in some continuous fashion based on a metric, such as the percentage vote that a candidate will receive. In this case, the market is predicting the expected value of the metric. Finally, the third type is spread betting, where participants bid on a cutoff that defines whether or not an event occurs. For example in sports, spread betting involves betting that a team will win by at least a certain minimum margin of points. This market predicts the median value of the cutoff.

As Berg, Forsythe, Nelson and Reitz found, some prediction markets like the Iowa Electronic Markets can be very accurate [2]. The market predicts the outcomes of political elections. Over four presidential elections, the market outperformed large-scale polling organizations to the point of having an error of only 1.5 percentage points a week before the elections compared to the 2.1 percentage points from the Gallup poll.

Other examples of prediction markets include Tradesports offering a contract on Saddam Hussein being ousted by the end of June 2003. The trading on this contract closely tracked both the expert journalists' assessment of the likelihood of war and oil prices. The Hollywood Stock Exchange (HSX), a prediction market on the box office success of movies as well as other related contracts, has also shown great overall

accuracy in predicting the actual revenue of movies. The market closes 4 weeks after release for most movies and 12 weeks for movies with limited release. This research focuses on making predictions on the HSX market.

For markets to succeed, they need a clear goal instead of vague contracts. For example, instead of "Weapons of Mass Destruction are not in Iraq," there would be contracts like "Weapons of Mass Destruction will be found in Iraq by [some specified date]". A market also needs clear motivation for the traders. There has been some discussion whether the use of play money vs. real money makes a significant difference. However, markets with accuracy such as HSX are examples that even trading for only play money and ego can produce an effective and efficient prediction market. Overall, prediction markets are valuable because they provide incentives for research, information discovery, and aggregating opinions with a market structure.

2.3 Prediction Movie Success

Movies provide an interesting and more controlled sandbox for prediction algorithms. Unlike many other domains, movies can be more easily compared to each other because they have some inherent normalization, i.e. they can be compared by success in the n^{th} week even if their release dates are different. While they too have many known and hidden factors, many significant factors affecting their success are more public and opinion-related than for example factors affecting the normal company. Unlike stock prediction, there are clear goals and a clear time line for movie success predictions. Also unlike stocks, the movies' success is much more directly affected by the general opinion and views of the populace, as these are the same people that go watch the movies and hence contribute directly to their success or lack thereof. The opinions of the random movie-goer can be shared with others and may actually affect whether other people will or will not go see the same movie based on the review.

2.3.1 Using Blogs

Sadikov, Parameswaran, and Venetis describe using blogs to predict movie success in their 2009 paper [13]. Since movies have a known release date, blogs theoretically provide a great medium to discuss and measure the hype surrounding a movie before and after it releases.

They filtered the top 300 movies from 2008 in terms of revenue and filtered out movies with common word names like "Wanted" that were likely to trigger false positives when computing reference counts in blogs. They generated 120 features with a basis in the following categories: movie reference counts in blogs, reference counts considering the ranking and in-degree of the blogs, features limited by a date range, features considering only positive posts, features addressing spam, and combinations of all of these. The reference counts considered factors such as whether the reference appeared in the title of the post. Their blog rankings weighted highly ranked blog references higher as well as references in blogs with higher in-degrees, a metric similar to page rank. Their date range analysis separated features by week from 5 weeks before release to 5 weeks after to try to capture the buzz each week. In regards to sentiment, they used LingPipe to perform hierarchical classification on the five sentences around the movie reference to determine positive posts in one approach and including negatives posts in a second. To address spam posts, they filtered on the length of the post primarily to avoid very short posts.

They evaluated their features using Pearson's correlation and Kullback-Leibler divergence against a few different outcome variables. These were average critics ratings, average user (viewers) ratings, 2008 gross sales, and weekly box office sales for the first five weeks. They had much more success predicting movie sales than the user or critic ratings. While the first week was best predicted by budget, they found correlations values as high as 0.86 between blog features and the weekly sales from weeks 2 to 5. References in blogs tended to precede movie sales by about a week, so they think predictions could be made up to a week in advance. Their study looked for correlations in the data but did not attempt to make any predictions.

2.3.2 Using the News

Having seen the predictive power of the media, Zhang and Skiena sought to investigate the effect of news on movie performance [19]. They wanted to show that using news data could be very informative, as commercially successful movies, actors, and directors all enjoy significant media exposure. The system Lydia was used to analyze the text of news articles. Compared to the Sadikov et al. study, they used movies spanning from 1960 to 2008 and worked with 498 movies in total. They also threw out movies with movie titles containing very common words that lead to false positives when using Lydia. The system allowed them to aggregate not only news reference counts but also positive and negative sentiment data on the movies from the news articles. They looked at news surrounding movie titles, directors, the top 3 actors and top 15 actors 1 week, 4 weeks, and 4 months before each movie's release date.

They built both regression models and k-Nearest-Neighbor (KNN) models. For the regression models, they tested for multilinear relationships between budget, holiday flag, MPAA rating, sequel flag, foreign flag, opening screens, and genres and the movie's gross income. They also tried building the model while leaving budget out. In addition to this, they tried using KNN under the assumption that similar movies will have similar grosses. Thus, they calculated the distances between movies based on their feature space values (the features are the same variables listed for regression) and determine which movies clump together into groups with similar traits. Movies in the testing set with similar traits as movies in the training set should theoretically perform similarly. While KNN with 1 neighbor performed poorly, they got good results using around 7 of the nearest neighbors.

Zhang and Skiena found that movie news references were highly correlated with movie grosses. Sentiment measures also correlated well. They found that models based purely on news could perform on par with models based on IMDb data, especially for high-grossing movies. Combining the news data with the IMDb data lead to the best results. Overall, regression worked better for low-grossing movies and KNN worked better for high-grossing movies. Finally, they found that article counts

worked well overall, but news sentiment only performed well in the KNN modeling approaches. Zhang and Skiena also claim their methods provide a big boost over previous work because they can make predictions before the movie release. Some models need up to a few weeks post-release to make accurate predictions. Being able to predict pre-release certainly makes the predictor more attractive.

2.3.3 Using Neural Networks and Other Models

Predicting movie success after release has already had some successful simple models. Litman and Ahn describe how most box-office receipts decrease after the opening week [8]. Finding that around 25% of revenue comes from the first two weeks, Sawhney and Eliashberg found that total box-office revenue can be forecasted with very high accuracy two weeks after release [14]. Seeking to improve on these models to provide more useful pre-release predictions, Sharda and Delen investigated applying neural networks to making pre-release predictions [15]. They used 834 movies from 1998 to 2002 with detailed data purchased from ShowBix Data, Inc. Their goal was not to predict exact revenues, but to classify movies into one of nine classes ranging from flop to blockbuster.

In their study, Sharda and Delen used seven independent variables that are well recognized from industry experts and previous studies as being effective. The seven are MPAA rating, competition from movies released at the same time, star value of the cast, content category/genre, technical effects, sequel status (is or isn't a sequel), and number of screens showing the movie when it opens.

They used a multilayer perceptron (MLP) neural network with two hidden layers for their model. Testing it with 10-fold cross validation, they classified success using average percent hit rate. There were two different hit rates: one for exactly correct classifications and one of classifications within 1 class of the correct classification. Over the five years of data they used, Sharda and Delen's exact classification hit rate averaged in the mid 30% range, but their 1-away hit rate was much more impressive, in the mid 70% range. In both cases, though, Sharda and Delen show that they were able to outperform recognized logistic regression, discriminant analysis, and

classification/regression tree approaches by four to five percentage points. Thus, they were able to demonstrate the effectiveness of the neural networks for movie prediction and project that their approach could be applied to other forecasting problems.

Alon et al. also investigated the use of neural networks for prediction, comparing the use of artificial neural networks, Winters exponential smoothing, Box-Jenkins ARIMA model and multilinear regression [1]. They analyzed the forecasting value of the four models on US aggregate retail sales over time, a data set that contains trend and seasonal patterns. Their findings showed that the artificial neural network faired the best overall, followed by Box-Jenkins, Winters, and finally multilinear regression. The artificial neural networks did the best in periods where economic conditions were relatively volatile while Box-Jenkins and Winters did the best under more stable conditions.

These findings also show that artificial neural networks should be considered as a viable modeling option. However, they are also difficult to train effectively including balancing the training, validation and testing sets, determining the correct number of starting nodes and initial weights and so on. Due to these concerns and some incompleteness in our data set, neural networks were not explored in our approach.

2.3.4 Using Social Network and Sentiment Analyses

Part of the hypothesis of the project is that social network position helps to predict movie success as discussed by Gloor et al. [6]. They generated social networks for the movies in three ways: using web searches, using blog searches, and using posters on movie forums. The network is built for example by Googling a relevant phrase such as "Slumdog Millionaire movie 2009," and then Googling for the pages that link to the top 10 hits on the first search. This can be done recursively a few times to generate a large graph where nodes are websites and edges are the links between websites as given from recursive searches for pages that link to the top ten results of the previous search. In addition to calculating the network importance of the movie title itself, Gloor et al. performed sentiment analysis on IMDb forums to gather the general mood towards a movie [6]. They used tags to identify references to the movie title or

shortened references thereof within a post. Lists of positive and negative words were then used to determine the general sentiment of the post towards the movie using standard information retrieval algorithms such as "term frequency-inverse document frequency." They also constructed a network using the post authors so that their betweenness centrality, i.e. social network importance, could be calculated. The positivity and negativity of a post were weighted using the betweenness of its author, thus weighting a more important poster's contributions more heavily in the overall sentiment score calculation.

Gloor et al. hypothesize that combining the sentiment towards a movie with its betweenness should in theory give a prediction about not only the general feeling about the movie, but also the magnitude of the feeling. Krauss et al. provided a validation of the predictive value of betweenness and sentiment in regards to film properties in their Oscar prediction paper [7]. They used the Oscar Buzz forum on IMDb and web/blog searches to predict which movies would win Oscars and which would perform well in the box office. Five of the seven movies ranked highly by their algorithm received Oscars, while another received a nomination and the last received nothing. They also found that movies with a high level of positive discussion performed well in the box office.

Chapter 3

Data Sources

To predict HSX prices, we gather many raw and derived independent variables. We categorize them as either Web Metrics, Social Network Analysis Metrics, or Sentiment Metrics. We also provide a justification of how these metrics help us capture the wisdom of the collective whole.

3.1 Web Metrics

We gather movie rating metrics from IMDb (www.IMDb.com) and Rotten Tomatoes (www.rottentomatoes.com) as well as box office performance data from Box Office Mojo (www.boxofficemojo.com). The movie trade volumes and quote prices themselves are gathered from the Hollywood Stock Exchange site (www.hsx.com). Part way through the project, HSX stopped publishing the trade volumes.

IMDb aggregates votes over a large number of users. Anyone can submit their rating of a movie to be included in the IMDb rating. Thus, IMDb provides a summary of the overall opinion of the collective whole and we hypothesize that IMDb represents the general feeling about the quality of a movie. However, since the voting is open, the rating is also susceptible to users trying to bias the vote artificially.

Rotten Tomatoes, on the other hand, collects the input of movie critics. We view this as an aggregation of the opinions of movie "experts" only. The number of contributors is smaller than IMDb, but in theory each vote may provide a better

quality input into the overall vote. Here we are polling the collective "expert" mind, but again we may find this vote susceptible to critic bias.

Both rating sites may also produce snow ball effects, especially on IMDb, where strong positive or negative reviews encourage more people to see or not see the movie. We hope to try to capture this behavior in our models if it does exist.

The Box Office Mojo provides day-to-day data on how much each movie has grossed in the box office to date. The income provides a reflection of the collective whole's general opinion regarding a movie. Initial high income that peters out represents a highly hyped movie that failed to live up to its reputation. Sustained income may represent the expected success of a good movie. Alternatively, a ramp in income over time represents a Black Swan such as *Slumdog Millionaire* that gains popularity as more people see and praise it. It also provides an empirical corrective factor for the final HSX price we are trying to predict.

3.2 Social Network Analysis Metrics

The first step to measuring a trend with dynamic social network analysis is the tracking of a movie title's relative importance on the Web, in the blogosphere, or in an online forum. We call each of these different online communication archives an information sphere. As an approximation for the relative importance of a concept in the information sphere, we calculate the betweenness centrality of this concept within the chosen information sphere. This means that we are extending the well-known concept of betweenness centrality of actors in social networks to semantic networks of concepts – movie titles in this case.

The betweenness centrality of a concept in a social network is an approximation of its influence on the discussion in general. Betweenness centrality in social network analysis tracks the number of geodesic paths through the entire network which pass through the node whose influence is measured. It is calculated as follows: find all the shortest paths between every pair of nodes in the graph. The betweenness centrality of some node A is the ratio of the number of these shortest paths A appears on (as

an intermediate node) to the total number of shortest paths. Thus, the value ranges from 0 to 1. In a star network structure, the center node will have a betweenness of 1 because every shortest path passes through it. The outer nodes will have a betweenness of 0. Brandes gives an explanation of the algorithm used to calculate this value efficiently [4]. Thus the higher the betweenness of a node, the more important it is because more shortest-path communications must go through it.

As access to knowledge and information flow are means to gain and hold on to power, the betweenness centrality of a concept within its semantic network is a direct indicator of its influence [17]. In other words, concepts with high betweenness centrality are acting as gatekeepers between different domains. While communication in online forums can be used to construct social networks among actors, we can also construct social networks from blogs and the Web. Although these semantic networks based on blog and Web links are not true social networks in the original sense, they are straightforward to construct by considering the Websites and blog posts as nodes and the links between the Websites and blog posts as ties of the social network.

Measuring the betweenness centrality of a concept permits us to track the importance of a concept on the Web or in the Blogosphere. This can be done either as a one-time measurement, or continuously in regular intervals over time, as Web pages, blog posts, and forum posts all have time stamps. We therefore periodically (e.g. once per day, once per hour, etc.) calculate the betweenness centrality of the concept. The resulting betweenness centrality is a numerical value between zero and one, with zero implying no importance of the concept in the information sphere and values above zero representing the relative importance in comparison to other concepts.

To build the semantic social network in an information sphere we introduce degree-of-separation search. Degree-of-separation search works by building a two-mode network map displaying the linking structure of a list of websites or blog posts returned in response to a search query or the links among posters responding to an original post in an online forum. For example, a search to get the betweenness of "Hillary Clinton" on the Web works as follows:

1. Start by entering the search string "Hillary Clinton" into a search engine.

2. Take the top N (N is a small number, for example 10), of websites returned to query "Hillary Clinton."

3. Get the top N websites pointing to each of the returned websites in step 2 by functionally executing a "link:URL" query, where URL is one of the top N websites returned in step 2. While the exact syntax and API varies between search engines, the meaning of the "link:" or equivalent query is to retrieve what the search engine considers to be "significant" websites linking back to the specified URL.

4. Get the top N websites pointing to each of the returned websites in step 3. Repeat step 4 up to the desired degree of separation from the original top N websites collected in step 2. Usually it is sufficient, however, to run step 4 just once.

Degree-of-separation search therefore is very similar to a domain-specific page-rank algorithm [5]. The betweenness metrics represent the general buzz on the movie from the web and from bloggers. We hypothesize that they will be useful variables because they are unconscious signals about a movie's popularity (or notoriety). That is, they are not calculated by active input from people and are therefore difficult to influence artificially.

3.3 Sentiment Analysis Metrics

To determine the general sentiment about the movies, we gather posts from IMDb forums. We previously used the following general forums: Oscar Buzz, Film General, Box Office, and Now Playing and Upcoming Films. However, we are also tracking communication on movie-specific forums to allow us to better differentiate which posts are about which movie. For this research, we only used one sentiment algorithm under development in our group to generate a positive and negative sentiment score. The group is also experimenting with other methods to improve the accuracy of the

sentiment algorithm. When using general forums, we also counted the occurrences of the movie's title, referred to as word count.

In addition to calculating a sentiment score for each post, we also build a social network of all the post authors in order to calculate their betweennesses. We then weight the post sentiment scores by the betweenness of its author. This gives the posts by more between, and we hypothesize more influential, authors relatively higher sentiment scores. Betweenness values range between 0 and 1. Authors are weighted using $1 + betweennesscentrality$, so that the default weight is 1 and more influential authors have slightly higher weights. These weighted scores give us the variables word count betweenness, positive betweenness, and negative betweenness.

Pang and Lee have shown that automatic extraction of words and word pairs leads to more precise results than manually selecting positive and negative words [12]. Our approach follows the basic "bag-of-words" approach which considers the co-occurrences of keywords in sentences or text [11]. A drawback of this approach is that it disregards grammatical dependencies in the analyzed data. This might lead to misleading interpretation in some cases. For example the statement, "Terminator is not good," would be classified as a positive sentiment with the simple bag-of-words approach that looked at single words only. In practice this problem seems to be rare, however. Matsuzawa and Fukuda state that 40% of analyzed keywords in the same sentence or text block show grammatical dependencies [10]. By reading a large sample of forum messages we empirically verified their finding that actors mostly use negative phrases rather than negating positive phrases when they wanted to express something negative. For example they use the phrase "is bad" instead of "is not good." We further reduce occurrence of this problem through not looking at the whole post but rather only words around a word anchor.

The starting point of the sentiment retrieval is the collection of word lists that constitute the initial bag-of-words. These lists were retrieved from the movie discussion on the IMDb forum and manually checked to assess the words' appropriateness. One list is used for positive words, one for negative. To deal with different cases, singular and plural forms, etc., we apply Porter Stemming [16]. Through the application of

Metric Categories			
Social Network Analysis, Sentiment, and HSX	Box Office	Rotten Tomatoes	IMDb
word count betweenness positive betweenness negative betweenness web betweenness blog betweenness HSX price (in millions) HSX trade volume	daily gross daily percent change weekly percent change number of theaters showing gross per theater gross to date	percent rating review count fresh review count rotten review count average rating	mean weighted average median total votes i-star votes (10 vars, i in $[1, 10]$) percent of votes for i-star (10 vars, i in $[1, 10]$)

Table 3.1: This table contains all the independent variables garnered from the data sources discussed in Chapter 3 and used as input for all the modeling approaches. Sometimes, the series of differences were also used, i.e. the series of the $i^{th} - (i-1)^{th}$ variable values with $i \geq 1$. These are denoted with the phrase "diffs" appended to the variable name.

stop lists to forum posts we sort out unimportant words like "the," "and," "or," etc.

When analyzing IMDb posts, we only consider the individual forum that discusses the current movie of interest. Each movie has its own message board which is being used as its document corpus. These words form the basis for the comparison with the bag-of-words. Generally, when more words from the positive list are present, the positive sentiment value is higher. The same is true for negative words and negative sentiment value. The sentiment algorithm basically counts the occurrences of words from the positive and negative bags-of-words in the posts and weights the counts with the author betweennesses. It should be noted that the analysis in Sections 4.1 and 4.2 used an older version of the sentiment algorithm which used the general forums mentioned previously and identified posts are pertaining to particular movies based on the presence of certain tags in the content. These tags were common variations of the titles of the movies aimed to capture the ways people typically referred to the movies in their forum posts. For example, *Terminator Salvation* could also be called "Terminator 2009" or "Terminator 4".

Chapter 4

Single Movie Models

This chapter describes modeling approaches that make predictions treating each movie individually. That is, we only look at current and historical data regarding the movie's performance over various metrics to make predictions on how well that same movie will perform in the future. We tested each model on many movies to gauge its overall accuracy.

In Section 4.1 we attempt to predict the direction of the price change on HSX, i.e. whether it will go up or go down on a given day in the future. Finally, Section 4.2 describes a linear regression approach to predict the magnitude of the daily price changes on HSX.

4.1 Price Change Direction Prediction

We briefly investigated predicting whether the HSX price tomorrow will go up or down based on the changes in variables on previous days. In the simplest case, we qualitatively tested the hypothesis that on day j, the sign of the delta of some independent variable between day $j-1$ and j would match the sign of the delta of the price between day j and $j+1$. As an extension of this idea, we looked at the sign of the delta over several independent variables to develop a majority vote for the direction of price change. For example, 5 variables with deltas of positive, negative, positive, positive and negative generate a 3-2 vote in favor of a positive price change

Independent-Variable Delta Sign	Price Change Delta Sign
sign(iv($j-3$)-iv($j-4$))	sign(pc($j-2$)-pc($j-3$))
sign(iv($j-2$)-iv($j-3$))	sign(pc($j-1$)-pc($j-2$))
sign(iv($j-1$)-iv($j-2$))	sign(pc(j)-pc($j-1$))

Table 4.1: We compare the sign of the independent variable (iv) delta of the three days previous to day j against the corresponding the price change (pc) deltas to see if the variable had been an accurate predictor in the recent past. Here the *predictionperiod* is equal to 3.

Independent-Variable Delta Sign	Price Change Delta Sign
sign(regression_line_slope([$j-6$,iv($j-6$)] , [$j-5$,iv($j-5$)] , [$j-4$,iv($j-4$)])	sign(pc($j-3$)-pc($j-4$))
sign(regression_line_slope([$j-5$,iv($j-5$)] , [$j-4$,iv($j-4$)] , [$j-3$,iv($j-3$)])	sign(pc($j-2$)-pc($j-3$))
sign(regression_line_slope([$j-4$,iv($j-4$)] , [$j-3$,iv($j-3$)] , [$j-2$,iv($j-2$)])	sign(pc($j-1$)-pc($j-2$))
sign(regression_line_slope([$j-3$,iv($j-3$)] , [$j-2$,iv($j-2$)] , [$j-1$,iv($j-1$)])	sign(pc(j)-pc($j-1$))

Table 4.2: We compare the *predictionperiod* signs of the regression line slopes, each computed from *trendperiod* independent variable (iv) data points, against the sign of the price change (pc) for the day after the last day of the *trendperiod*. Here the *predictionperiod* is equal to 4 as we have 4 rows of sign match ups. The *trendperiod* is 3, as the regression line slopes are computed using 3 ordered pairs of [*daynumber*,*ivvalueonthatday*].

as our final prediction. The next question is to choose which 5 (or n) variables to use. For this, we evaluated each independent variable over *predictionperiod* days immediately before day $j-1$ to check how often its delta prediction was correct during that period. Table 4.1 gives a specific example with a *predictionperiod* of 3 days to show which deltas are compared. We then choose the n variables with the highest ratios of correct predictions over the *predictionperiod* days, i.e. the highest recent historical accuracy, as the voting variables for the final prediction.

Expanding from using the sign of the delta, which is the sign of the slope between the two consecutive points, we also tried the slope of the regression line incorporating more data points. Instead of the consecutive deltas as shown in the first column of Table 4.1, we calculated the regression line of every *trendperiod* consecutive days and made the line's slope sign our prediction instead. Table 4.2 gives an example of the points used, again assuming *predictionperiod* is 3 days. The *predictionperiod* days were used to calculate correct ratios as before. The ordering and voting with variables is also the same as before.

We theorize that future behavior reflects recent behavior, i.e. that the future is a

product of the past. We aim to find independent variables whose behavior matches the behavior of the price in terms of ups and downs. For example, it is possible that if the positive sentiment metric goes down today, the price will go down tomorrow. Table 3.1 lists all the independent variables we considered, including the price itself. In the latter case, we use price changes in the past to predict price changes in the future. In addition to variables in Table 3.1, we also derive a second set of independent variables composed of the differences of consecutive days' values of the independent variables found in the table. The regression line approach refines over the delta to provide a smoothing effect over sudden changes and oscillations. The line provides the direction of the overall trend in the price going up or down. To diversify our prediction and make it more robust, we use the n most accurately predictive independent variables to date for each subsequent prediction to protect against occasional aberrations in individual variables.

4.2 Price Change Magnitude Prediction Using Linear Regression

For our first attempt to predict the magnitude of the movie stock price, we used linear regression to predict the price tomorrow on day $j+1$. We based our predictions on a model slope and intercept built using previous independent variable and movie stock price data. Over many trials, we varied the *trendperiod*, again the number of days of data used to build the model, and the *predictiondays*, or the number of days the price was shifted forward for each independent variable and price value pairing. To clarify, to test the hypothesis that day j's independent variable value affects day $j+1$'s price and verify that the previous 5 days data best helps us predict the pattern for tomorrow, *predictiondays* would be set to 1 and *trendperiod* would be set to 5. To calculate the prediction for day 7, we would use the independent variable from days 1-5 and the prices from days 2-6 to build a linear regression model. Then we enter the independent variable value at day 6 as input to compute the predicted price

output for day 7. Increasing *predictiondays* to 2 would mean that the independent variable from days 0-4 would be used with the prices from days 2-6 to build the model. Thus, *predictiondays* represents the lag in the effects of the independent variable on the price, while *trendperiod* encapsulates the number of the days that define a short term trend to model in the price. Especially when considering movies, it is very likely that there will be a lag between a positive or negative set of reviews posted online regarding a movie and the subsequent boosting or damaging effect on its price and revenue. We would also expect that *trendperiod* would remain relatively short to capture changing viewer responses.

We tried many combinations of values for both variables over all movies to analyze how much the price changes lagged in changes to each independent variable and how many days were typically required to build a model to capture the current trend in the price. For each value of *trendperiod* ranging from the minimum 2 days to 14 and *predictiondays* ranging from the minimum 1 day to 14, we cycled through all our historical data for a movie and built a model for every sliding window of *trendperiod* days to make a prediction. This means for $trendperiod = 5$ and $predictiondays = 2$, we would make a total of 30 predictions if we had 37 days of data.

4.3 Single Movie Model Results

The price change direction predictions for Section 4.1 only suggested that certain variables seemed to track the price changes. The approach was tried on movies with a trading volume averaging over 1 million per day to ensure we only used actively traded movies. Table 4.3 shows the independent variables that were most often used to predict the price change direction. These are the variables that were most often found to track the changes in price. The table suggests that the HSX price is the best predictor for the next price change. This is not surprising since prices usually do not oscillate wildly but instead have a general trend of going up or down for several consecutive days. Also of note are the other most used predictors are the ones we hypothesis will be helpful predictors, i.e. the sentiment and betweenness variables.

Independent Variable	Usage Frequency
HSX price	248
word count betweenness	161
wordcount betweenness diffs	143
HSX price diffs	114
positivity betweenness	101
negativity betweenness diffs	96
negativity betweenness	95
positivity betweenness diffs	91
blog betweenness	86
blog betweenness diffs	85
web betweenness	84
HSX trade volume diffs	71
HSX trade volume	55
web betweenness diffs	44
daily gross	32
gross per theater	32
gross per theater diffs	29
daily gross diffs	27
daily (gross) percent change	25
number of theaters showing	24
gross to date diffs	22
number of theaters showing diffs	17
gross to date	16
daily (gross) percent change diffs	13
rotten tomatoes percent rating	9

Table 4.3: This table contains the frequencies with which each independent variable was used by the price change direction predictor when drawing a consensus vote to generate its predictions. The top 25 are shown here. Of note is that this list is dominated by the HSX price and then sentiment and betweenness variables.

However, these results are a little misleading for the same reason that the price seems to be a good predictor. The coarse-grained nature of this approach with only 2 values for each outcome leaves it susceptible to coincidental correlations. Furthermore, simply knowing if a price will go up and down is not very useful without providing a prediction for the magnitude of the change. Without magnitude, it is difficult to make meaningful decisions about which movie stocks to buy or sell each day. Also, the coarse and local nature of this type of prediction does not allow us to guess at the actual final stock value. While not necessarily directly useful, these results suggest that these variables may make good indicators when we try to predict the final stock price using different methods.

When predicting the magnitude in price change using linear regression, we com-

pared the predictions against the actual prices for those days to evaluate each pair of (*trendperiod*, *predictiondays*) variable values. Specifically, we computed the mean error, standard deviation of error, and mean squared error. Since each prediction also has a correlation, we also computed the mean correlation and correlation standard deviation. Some the results here have reasonably low error, indicating we may be able to predict day to day changes reasonably to make trades. Table 4.4 shows results for the movie *Up*. The number of theaters showing the movie each day was the best predictor for the next day's movie stock price with an 8 day delay representing the lag between people seeing the movie and reacting to it on the prediction market. A mean squared error of 10.387 accumulated over 10 predictions in this case seems reasonable (also recall all HSX prices are in millions). For *Up* at least, we found most of the best results using IMDb data. The IMDb-based predictors tended to have short *trendperiod*s. In conjunction with the longer *trendperiod* for the number of theaters showing the movie, we get a logical progression of people seeing the movie, then IMDb receiving votes a few days later, and finally HSX reflecting these events in its price a few more days later. We see that the volume of middle-ground votes for 3, 5 and 6 do best in predicting the day-to-day HSX rather than the extreme votes. Perhaps these votes provide a better look at the general feeling of the public rather than the fans and the haters. From this, we can interpret that the movement of HSX prices tends to follow the opinion of collective general public presenting as captured via voting on IMDb. This provides another confirmation that the collective whole can give us insights into the prediction market's behavior.

It should be noted that the mean squared error is calculated using the error between each predicted price and actual price for each day. The correlation coefficient mean is an average of the correlation coefficients of all the regression models used for the predictions. Thus, the accuracy of the predictions and the correlation coefficients in this table are related, but not directly linked. We have not yet developed any trading strategies to apply these results.

IV	Trend Period	Prediction Days	Mean Error	Error Standard Deviation	Mean Squared Error	Correlation Coefficient Mean	Correlation Coefficient Standard Deviation
number of theaters showing	8	2	0.006315	3.380257	10.387	0.765246	0.1445
6-star IMDb votes	2	5	.136708	3.348881	10.533	.428571	.937614
percent of 3-star IMDb votes	8	14	-1.366	3.134408	10.989	-0.46962	0.521188
3-star IMDb votes	3	11	-1.575	3.155722	11.609	-0.652	0.500621
5-star IMDb votes	2	4	-0.2703	3.530512	11.759	0.428571	0.937614
percent of 3-star IMDb votes	9	14	-1.93602	2.972594	11.905	-0.50907	0.519679
IMDb weighted average	3	12	0.1785	3.685474	12.709	-0.29391	0.756806
6-star IMDb votes	2	8	-0.97831	3.556334	12.761	-0.07692	1.037749
IMDb mean	4	11	-1.52655	3.359891	12.813	-0.63204	0.356732
5-star IMDb votes	2	8	-0.17524	3.706572	12.853	0.384615	0.960769

Table 4.4: This table contains the top 10 linear regression price change magnitude prediction results for the movie *Up* ordered by mean squared error. IMDb was largely one of the best predictors. The first column contains the name of the independent variable (IV).

Chapter 5

Multiple Movie Models

This chapter describes modeling approaches that make predictions by combining the metrics of many movies together to build a more general model about movie behavior rather than building models specific to each movie as in Chapter 4. To test the models, the leave-one-out (LOO) strategy is employed. Under this testing strategy, we build the model using all the movies with sufficient data sets except one and then use the model to predict or classify using the movie we left out. We repeat this approach n times for n movies, leaving each out in turn. Error results are averaged to evaluate the overall effectiveness of the modeling approach. This approach provides statistically accurate results because we assume each movie to be an independent and identically distributed (IID) random variable. Our model approach simplifies the real-world problem by assuming there is no interdependence between movies that affects their performance in the box office. Thus, the order in which movies came out makes no difference to this modeling approach. Both models in this section seek to make longer term predictions. They used data from two weeks prior to release date (as given by Box Office Mojo) through a week post-release to make their predictions about each movie's final box office revenue. Using this approach allowed us to compare movies that came out on different dates more directly.

In Section 5.1, we develop a classifier to place movies into three classes regarding their degree of success or lack thereof. Section 5.2 describes correlations between independent variables on a given day and gross revenue of each movie. Subsections de-

scribe linear models based on this approach that make early and effective predictions of each movie's actual revenue figure.

5.1 Classifying Movies By Degree of Success Using Bayes Rule

We categorized the success of movies based on the ratio of their gross income to production budget using three groups. Group I movies had a ratio of less than 1. These movies were flops that did not even recoup their investment. Group II comprised movies with a ratio between 1 and 2, i.e. movies that did decently. Group III held the most successful blockbuster movies making at least double what was invested or more.

We first classified each of the 30 movies with sufficient daily betweenness and sentiment values for the time period two weeks before release through a week after release into Groups I, II, and III based on their revenue to production budget ratios. Then we calculated the product of blog betweenness, or buzz, and positive and negative sentiment value for each day and summed the products together. Here positive sentiment values were positive numbers and negative sentiment values were negative numbers. The net sum represents the overall feeling including some buzz about the movie a week after release. We then took the absolute value and log of the sums to turn them into strength-of-feeling (whether positive or negative) scores of comparable scale.

The hypothesis for this model is that these scores are normally distributed within each of the three categories. We computed the mean and standard deviation of the strength-of-feeling scores for each category. Recall by Bayes rule, if g is the group for movie x, then $P(g|x) = P(x|g)P(g)/P(x)$, which is proportional to $P(x|g)P(g)$ since $P(x)$ does not depend on g. At this point, we did not have substantial data to suppose our prior $P(g)$ was not uniform between Groups I, II, and III. This could be refined in the future. Thus, $P(g|x)$ is proportional to $P(x|g)$, which we calculated using the

normal distribution probability density function for which x, the group mean, and the group standard deviation were inputs. We attributed x to whichever group gave the highest value for $P(x|g)$. Using LOO, we generated this classifier and attempted to correctly reclassify the left-out movie.

5.2 Correlations and Linear Models Over All Movies

In this section, we consider all our movies over the specified time period and generated correlations between each independent variable and the two different dependent variables: the HSX movie stock delist price and the final gross revenue of the movies. Since movies are delisted four weeks (or twelve weeks for limited release movies) after release on HSX, the final gross revenues are higher and represent the actual earnings.

Subsection 5.2.1 explains the correlations and linear predictors for single variables. Subsection 5.2.2 explains how we got improvements combining the best independent variables from Subsection 5.2.1.

As a prediction market, HSX is designed with the fundamental goal of predicting box office success correctly and lays down a baseline for comparison. We compared our models' errors against using the HSX prices themselves as predictions to evaluate the degree of our success.

5.2.1 Single Variable Correlations and Linear Models

For each independent variable for each day in the time period, we computed the correlation between the independent variable for each movie and each of our two dependent variables. For example, on release day using the independent variable HSX price, we found the correlation between the series of HSX prices that day against the series of HSX delist prices for all movies. We did the same for the series of HSX prices on release day against the series of box office gross revenues. Since our time period spans 20 days, we had 20 correlations for each dependent variable. The number of data points in each correlation varied with the independent variable and the day in question because the data set was not populated for all movies for all variables on

every day.

We also built a single-variable linear regression model for each independent variable for each day against each of the two dependent variables. Here, the LOO strategy was employed to calculate the mean squared error of using each independent variable as a predictor for the final HSX delist price and in a separate model as a predictor of the final box office gross.

5.2.2 Multiple Variable Correlations and Linear Models

In addition to testing each variable individually for its predictive value, we tested linear combinations of seven variables. Five of the seven were the ones with the five best single-variable correlations: HSX price, Rotten Tomatoes review count, 1-star IMDb votes, 10-star IMDb votes, and total IMDb votes. We also used the web and blog betweenness variables which we conjectured would provide incremental refinements. For this approach, we did not have sentiment data to incorporate as well. In this case, we only had one correlation and error calculation per day per dependent variable.

5.3 Multiple Movie Model Results

Classifying movies with normal-distribution-based Bayes classifier, we initially found the classifications to be correct 53% of the time, which is better than 33% for random guessing. Table 5.1 shows how often movies in each group were categorized correctly. Group I and III movies were classified pretty accurately, but Group II movies had many misclassifications. Table 5.2 shows how often a movie classified as being in a group was actually in that group. It should be noted that Group II movies showed very high success here at a 100%, but only 2 of 12 movies were identified as being in Group II. Finally, Table 5.3 shows perhaps the strongest result: A movie classified as Group III was really unlikely to turn out to be in Group I and a movie identified as in Group I was really unlikely to be in Group III. Thus a flop suspect identified by our approach rarely made it big and a movie we named as a potential blockbuster

Group	Size	Percentage of Movies Classified Correctly Into This Group
Group I	11	72.7%
Group II	12	16.7%
Group III	7	85.7%

Table 5.1: Success rates of classifying movies into the correct groups using the normal-distribution-based Bayes classifier.

Group Movie was Classified as Being In	Percentage of Those Movies Actually In That Group
Group I	56.5%
Group II	100%
Group III	55.4%

Table 5.2: These percentages compare the number of movies classified correctly against false positives.

rarely flopped.

From this we see that the classifier certainly has some noise from this small data sample, but that there do seem to be distinguishing characteristics between the groups. While confirming total misclassifications are rare is not a huge result in itself, we think these results help show that there is promise in pursuing this type of classification further.

From this we see that the classifier certainly has some noise from this small data sample, but that there do seem to be distinguishing characteristics between the groups. Indeed, Figure 5-1 shows the average sentiment score for the movies in each group over time. Before release day at day 0, the Group I films, or flops, have the lowest profile and least hype. After release, they receive some negative feelings, followed by an upsurge and a big crash before another recovery. We can explain this behavior as little hype, followed by some excitement hoping the movie will do well, followed by people who saw it in the first week realizing they did not care for it and finally some people appreciating it after all. In theory, the people seeing it later are the ones who are either interested or with less strong emotions as the excited

Percentage of Movies Classified as Group III, But Actually In Group II or III	82.4%
Percentage of Movies Classified as Group I, But Actually In Group I or II	88.9%

Table 5.3: This table shows that movies identified as successes usually did OK or quite well, but rarely actually ended up doing poorly. Likewise, movies classified as flops were either flops or did OK, but rarely became successes.

opening-weekend viewers, which might explain the second upsurge. Meanwhile, the Group II films, or so-so films, get a little hype before the release with speculation. As they come out, they start strong until people realize they were not so great after all and then they fizzle out in terms of receiving strong positive or negative sentiments. Finally, the Group III films, or blockbusters, start with a little hype before release, and then start strong on opening weekend and stay strong. There are some small dips with people not as impressed as they thought they would be, but overall the sentiment directed towards the movie is very positive. Perhaps after the first week of release, day 6, people who saw it early on have finished commenting and the next wave of weekend-movie go-ers are still getting ready to go see the film, explaining why the daily sentiment is not as high as before. Expanding the model data set to additional days/weeks would perhaps help confirm this. While confirming total misclassifications are rare is not a huge result in itself, we think these results help show that there is promise in pursuing this type of classification further.

The best correlations from Section 5.2 are shown in Figures 5-2 and 5-3. Only the best five models as well as the blog and web betweenness models are discussed in terms of correlation and predictive accuracy. They are named after the independent variable used to build them. The correlations for the multilinear model for each day are shown as well.

In Figure 5-2, we found a few key observations. The best single-variable correlating variable was the HSX Price itself, showing that the prediction market fulfills its role effectively. The non-trivial correlations on R^2 between .4 and .6 for the Rotten Tomatoes and IMDb variables are also noteworthy. We found that the sheer number of votes on Rotten Tomatoes, rather than the number of fresh (positive) or rotten (negative) votes specifically, had the best correlation to the HSX delist price. In other words, movies that more critics bothered to review performed better overall, suggesting that more publicity of any sort is better than less publicity.

The total IMDb votes showed a very similar pattern. We saw that the greater interest of the general public instead of critics in this case, the fact that more people were voting on movies (and hence probably seeing them in the general case), may have

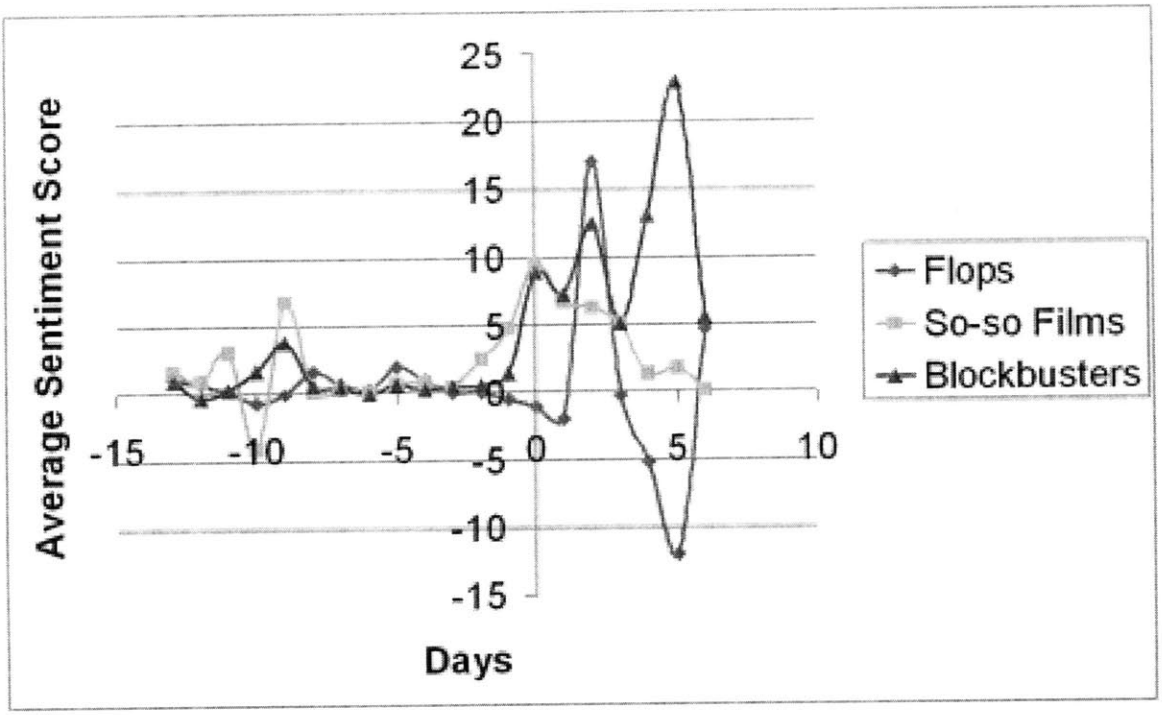

Figure 5-1: This figure shows the daily average sentiment score for the movies in each group. Day 0 is the release day. The range is two weeks before release to a week after. The flops from Group I start quiet, get a negative divot, generate a hopeful spike, and then crash with negative comments. The so-so films of Group II start with some comments before release, then start well when they open, but peter out as people realize that they are just so-so. There are neither strong negative nor positive sentiments about these films. Finally, blockbuster, or Group III, movies get some early hype, start strong, and mostly stay strong with slight dips in positivity, perhaps due to under-met but still satisfied expectations. The dips could be a little backlash from those who had very high expectations.

led to the movie performing better on HSX. It is also interesting that the number of 1-star and 10-star votes correlated well with the HSX delist prices. This also suggests that number of people with strong opinions on a movie correlates well to its success, whether those opinions are strongly positive or negative. Movies with fewer people with strong opinions performed less well. We found the HSX delist prices not to be correlated with the web and blog betweennesses themselves.

Apart from a dip in correlation which corresponds to some aberration in the 1-star and 10-star votes as well, the multilinear correlation was the strongest. Using a multilinear model, we were able to refine the high correlation of the HSX price with

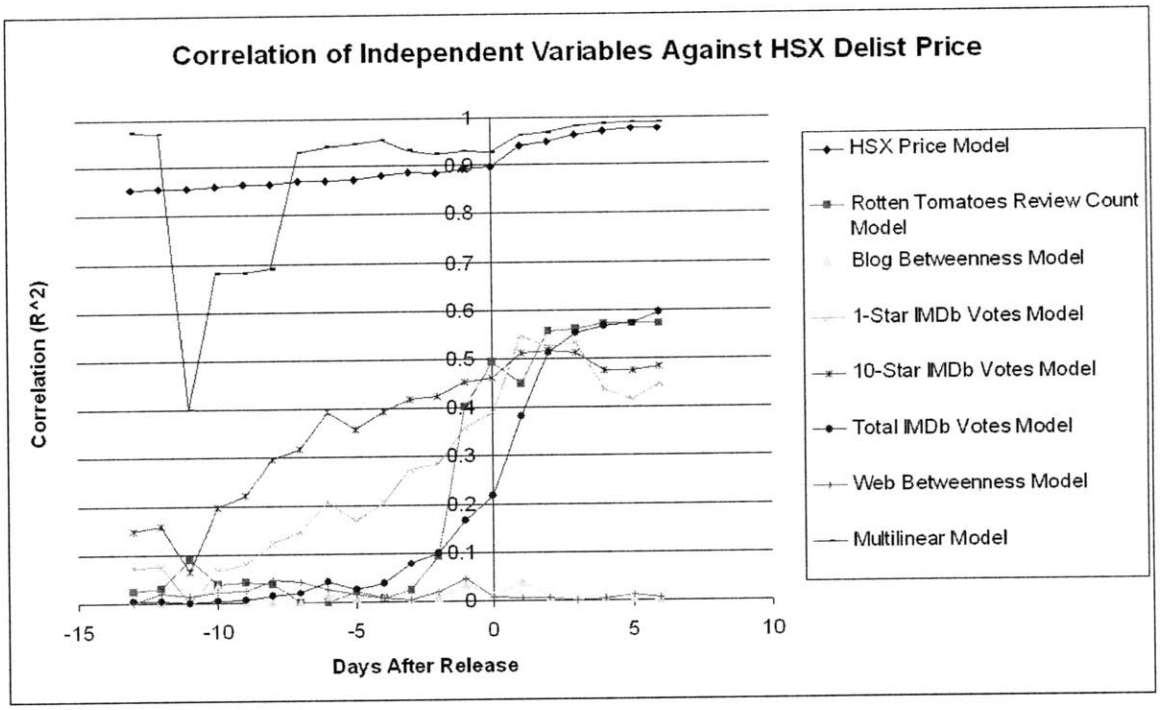

Figure 5-2: This figure shows the correlation of our independent variables against the HSX delist price of each movie. The HSX price itself even two weeks before release had a high correlation to the HSX delist price four weeks post-release. Only the multilinear model combining the HSX price with the other independent performed slightly better. The Rotten Tomatoes and IMDb-based variables performed fairly well, showing the effectiveness of the wisdom of the crowds. Finally, the betweenness variables on their own showed next to no correlation.

the other independent variables to find a consistently stronger linear correlation from a week before release forward.

The HSX price and multilinear model generally improve over time as the release date nears and passes. The Rotten Tomatoes review count correlation jumps just before release date, which must be when many critics see the movie early and report their feedback. The total IMDb votes correlation jumps just after the release date, after the first wave of the public has a chance to see the movie and also post its opinion online. We conjecture that the higher correlations before release of especially the 10-star votes results from people with strong opinions finding ways to see or learn about a movie before it comes out, perhaps through sneak previews. A movie that does well is more likely to have had more fans with strong feelings who made efforts to watch or see the movie ahead of time and voiced their opinions – which may also

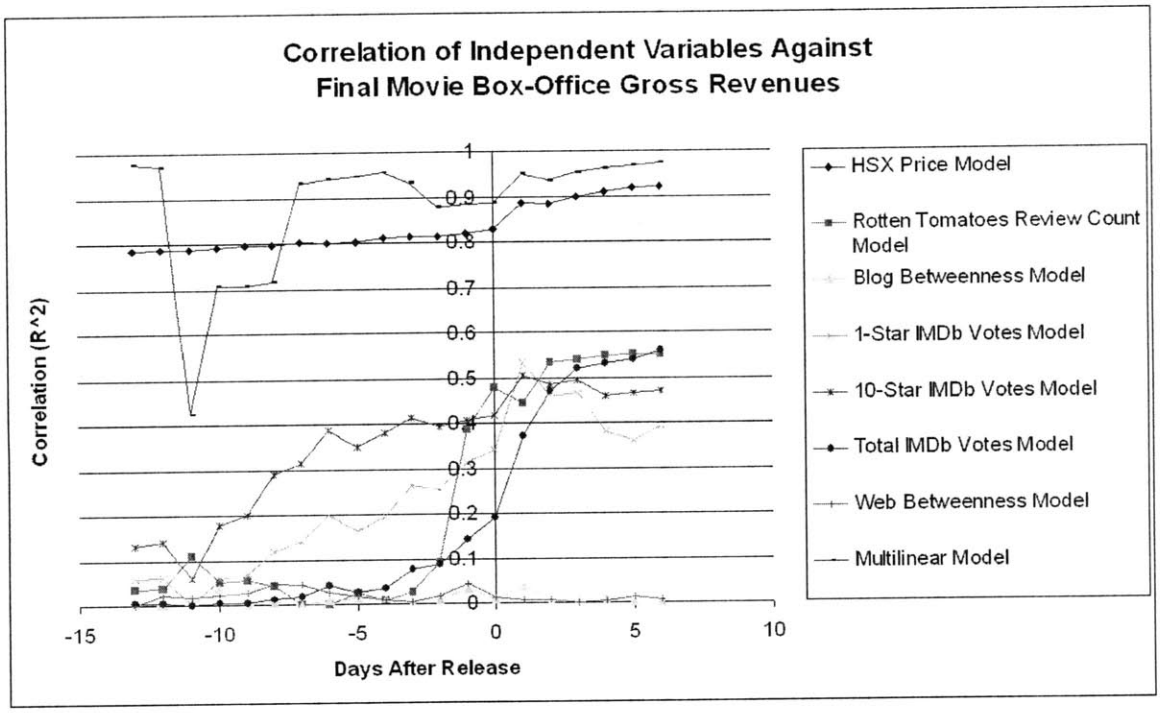

Figure 5-3: This figure shows the correlation between of our independent variables against the final movie box-office gross revenue of each movie. The overall patterns are largely very similar to those seen in Figure 5-2, but the gap between the HSX price alone compared to the multilinear model is slightly larger in this case.

help encourage others to see the movie and boost its overall performance.

We see very similar behavior and can make similar conclusions about the correlations with final movie box-office gross revenues shown in Figure 5-3. The biggest difference is that the correlation of the HSX price model is slightly lower and the gap between that model and the multilinear model is larger. The slightly poorer correlation between the HSX price and the final gross revenue is understandable since the HSX market aims to predict the delist price, a figure that is usually the box-office revenue after 4 weeks and hence lower than the final gross revenue. That said, the correlation is still quite strong. Thus, it appears that the movies generally seem to grow their revenues at the same rate before and after they are delisted from HSX. The multilinear model is able to refine the HSX price with the other variables to better correlate to a value that the HSX price itself is not originally meant to predict.

Seeing the high correlations from some of the independent variables, we then built and tested single-variable and multi-variable linear regression models. Figures 5-4 and

5-5 present the mean squared error from each of the models based on the independent variables used for the correlations. The figures also include the error from using the HSX price itself as a predictor.

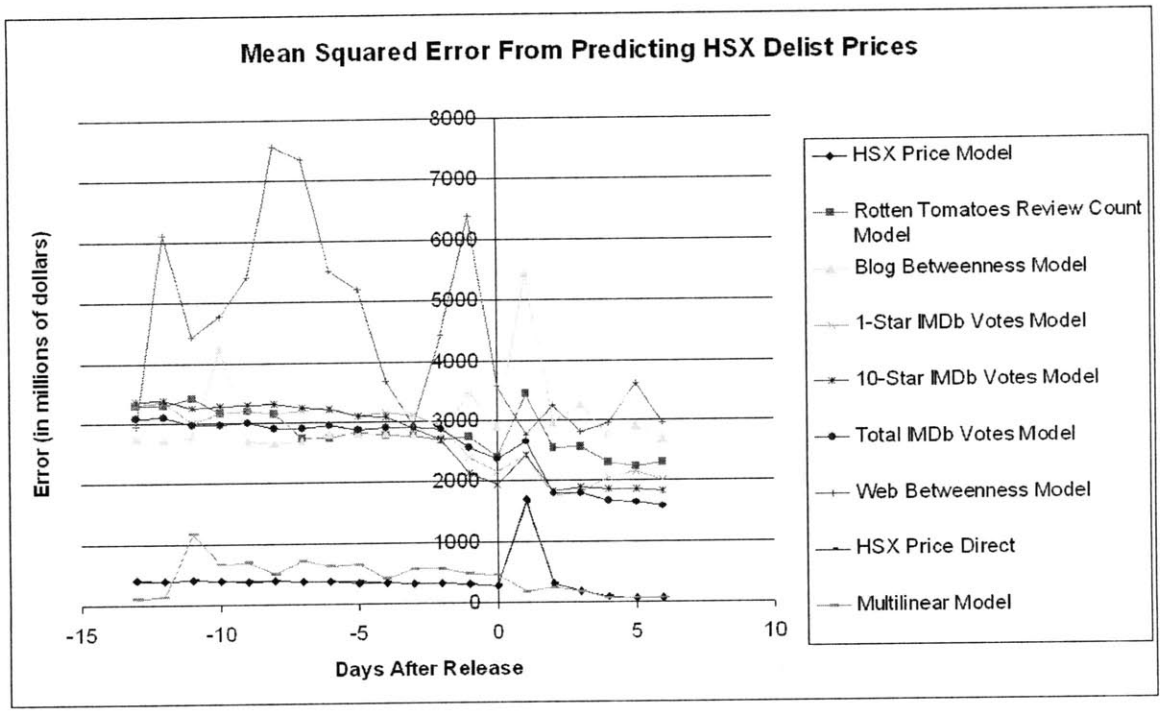

Figure 5-4: This figure shows the mean squared error from predicting the HSX delist price using a linear regression model with each of our independent variables. The error from the variables with the best correlations are quite low, especially a few days after release where the best errors are between 56 to 67. As expected, the errors from the Rotten Tomatoes and IMDb variables are higher and the betweenness ones are the worst. Of note is that the multilinear model is robust to the spike after release in error that the other single-variable models all experience.

In Figure 5-4, the errors of most models are fairly high. However, the linear-regression prediction model built using the HSX price, the HSX price itself, and the multilinear-regression prediction model all show low and comparable errors. Before release, the HSX price model and HSX price itself show very similar and low errors, though the HSX price itself is almost always slightly better. The multilinear model possesses higher error before release, likely because most of its contributing variables have high errors before release too. We see a definite improvement after release as the variables begin to more accurately reflect viewers' opinions on the movies instead of their expectations about the movie. In fact, the multilinear model becomes the best

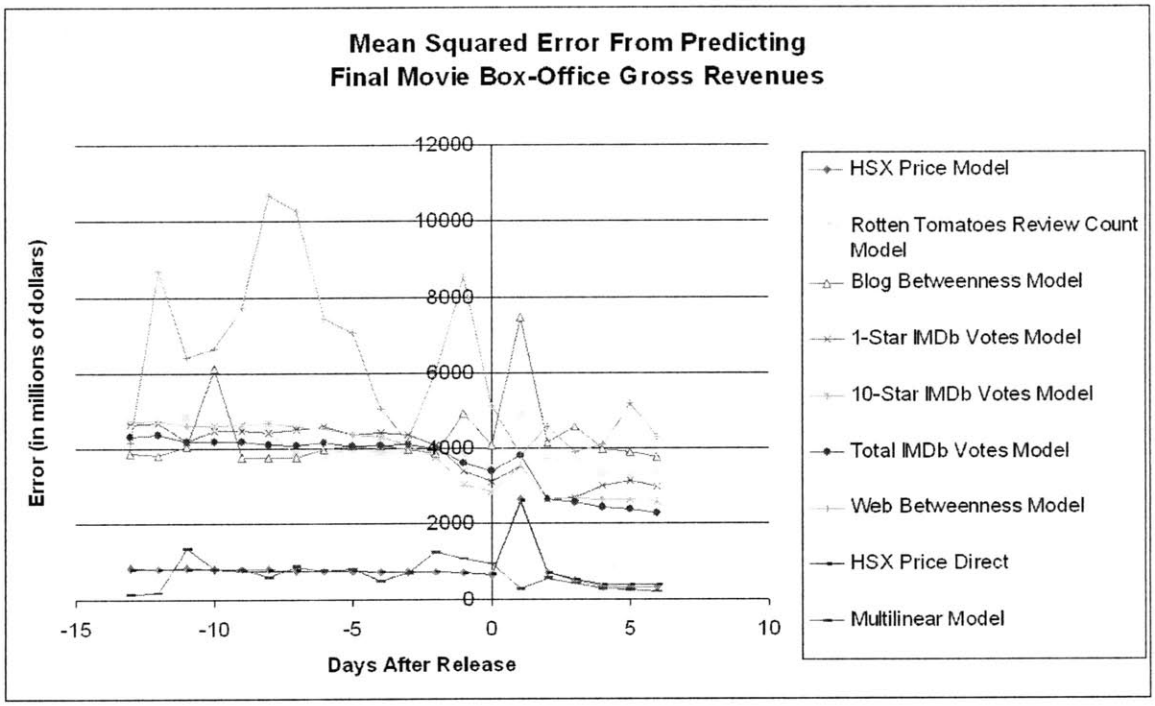

Figure 5-5: This figure shows the mean squared error from predicting the final movie box-office gross revenues using models generated from our independent variables. Again the results mirror those in Figure 5-4, though the error magnitudes are higher. This is to be expected since we are using predictors like HSX prices which are meant to predict HSX delist price to predict the final return instead. The effectiveness of the HSX price models and the multilinear models shows how HSX prices can be used quite effectively to predict the final revenue even if that is not the direct goal of the prediction market. The best errors are between 213 and the mid 300's.

after release. The HSX price model and HSX price see a spike in error immediately after release. This is likely from the gut hopeful or cynical reaction of HSX traders after they see the movie on opening day. The price settles back after a day as people consider their decisions. By 6 days after release, the three approaches have mean squared errors of less than 67 and the multilinear model only beats the HSX price model by about 10, or 15.18%,and the HSX price itself by about 5, or 8.66%. For the HSX price model, the slope of the linear regression model starts around .95 13 days before release and reaches .999 by 6 days after release. This shows again that the HSX price itself basically is a great predictor of the movies HSX delist price, as can be expected. The robustness of the multilinear model after release to the spike in the HSX price shows that including other variables improves the strength of the model.

Though we found the multilinear model to be better than the HSX price model and HSX price itself, the difference is not very large.

As before, we see similar general patterns in Figure 5-5. However, the magnitudes of the mean squared error are much higher. Here we must keep in mind that we are trying to predict a value that the HSX market is not designed explicitly to predict. That said, the final box-office gross revenue is not completely unrelated to the HSX price either. Given the high correlation in Figure 5-3, the relatively low error in Figure 5-5, and the slope of the HSX price model ranging between between 1.06 before release to 1.15 after release, we find that HSX quite consistently underestimates the final gross revenue. The underestimating is to be expected, but the consistency of the degree of underestimating over all movies is fascinating. The HSX price model now of course outperforms the HSX price itself, which is not meant to estimate the final gross revenue. Now that we are predicting a value that none of the independent variables is singly explicitly meant to predict itself, the multilinear regression shows a more significant refinement over the variables individually by combining the useful contributions of each into a single estimate. This time, the mean squared error improvements 5 and 6 days after release after about 100, or 32.00%, over the HSX price model and about 165, or 43.72% over the HSX price itself. Figure 5-6 shows the three best predictors alone to reduce the range of the error scale and demonstrate the superiority of the multilinear model after release.

Furthermore, we observed that the mean squared error is often skewed by a few very inaccurate estimates rather than a general model failure. Thus, in most cases, the best models produce quite accurate predictions. Removing these aberrations would further reduce the model's mean squared error, but it is difficult to identify which movies will be the problem movies. However, the fact that the HSX price is a strong proxy of the HSX delist price means that on any given day, we have the current values of the other independent variable and effectively also have the eventual value of the HSX delist price, in the form of the HSX price that day. Section 6.2.2 will discuss some suggestions to use this fact to uncover aberrant movies.

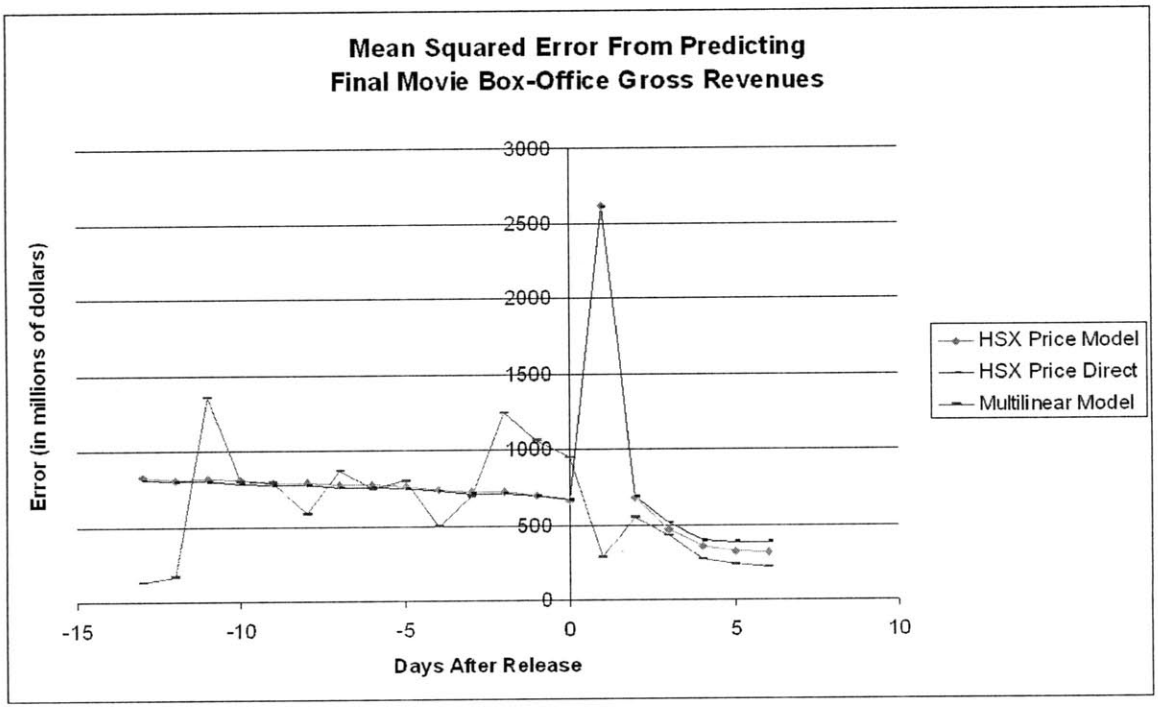

Figure 5-6: This figure shows a closer look at the error from the best variables predicting the final movie box-office gross revenues. Again the multilinear model is robust to the spike post opening day. The figure also more clearly demonstrates the relative differences between the three best models and we can see how the multilinear model beats the HSX price model by 100 and the HSX price directly by 165.

Chapter 6

Future Work

There are many ways to build on and improve the prediction models presented in the findings here. These are a few ideas for possible next steps in continuing this research.

6.1 Betweenness Centrality

The goal of the betweenness centrality calculations is to capture the general feeling about a query term, in our case the movies, on the web or blogosphere. In this research, the web and blog betweenness centralities were calculated using a search engine query of the form "*movie_name* movie 2009". We hypothesize that this query gives us the general buzz associated with the movie. However, we found little differentiation between the betweennesses of different movies. We conjecture that altering the query to include a descriptive adjective or biasing term will provide us with topic-specific buzz that may give us more differentiation between movies. Thus, one avenue of future research would be to investigate using queries such as "*movie_name* boring 2009" or "*movie_name* success 2009".

6.2 Multiple-Movie Model Correlations

6.2.1 Improving the Model

We found a quite strong linear relationship between the independent variables mentioned in Chapter 5, Section 5.2 and the dependent variables of HSX delist price and final box-office revenue. Thus, we do not think another type of model is necessarily required. However, additional independent variables could be investigated to refine the model, including combinations of existing variables such as the product of betweenness (or buzz) with other independent variables to provide a buzz-weighting factor.

6.2.2 Finding Movies that Aberrate from the Model

We suggested previously that a large proportion of the predictive mean squared error came from movies that strongly aberrated from the model. Here we discuss one potential approach to seek out these movies explicitly. Finding aberrating movies early and thus identifying very successful or unsuccessful movies ahead of time would be a very useful result.

In Section 5.3, we found that HSX price proved to be a strong proxy for the eventual HSX delist price nearly two weeks before release. Thus, in addition to the historical data set, we then theoretically have the independent and dependent variable values for the movie we are trying to predict for. If the current movie's data drops the correlation or creates high error in the prediction model, this could be a suggestion that it will not follow the normal pattern exhibited by most movies and be significantly more (or less) successful. Regrettably, there was not sufficient time to test this hypothesis for this thesis. A good starting step to check this theory would be to see if the movies in the historical data set that do not fit well into the multilinear model also have noticeable characteristics such as high (or low) revenue to production budget ratios or other signs of great (or minimal) success.

6.3 Sentiment

This thesis did not focus too greatly on the specific sentiment algorithms. Improving these would help provide a better metric of the public feeling about a movie, which theoretically should correlate to the movie's success. After all, movies people think well of should do well – so we need a good way to know people think well of a movie. The fairly successful multilinear model from Section 5.2.2 of Chapter 5 did not use any sentiment inputs. We conjecture it would become even stronger with accurate sentiment metrics.

6.4 Time Effects

None of the models here consider time-related effects. Movie producers time releases around weekends, holidays, and other big events, including the release of other competing movies. Differing opening day successes may be explained better by the timing of such external events rather than just the quality of the movie itself. A more complex model may be able to consider these factors when comparing the success of two different movies.

6.5 Additional Variables

Along with same of the variables analyzed in this thesis, there are many endogenous variables that have been shown to predict well and should be incorporated. These include variables such as movie budget, actor star power, and movie rating.

6.6 Trading

Now that we have built a few predition models, it would be interesting to develop a trading strategy to trade on HSX based on the models to compare which ones perform the best against each other, against not trading, and against random trading. Basically, the goal is to see if the prediction models can be used to improve a trading

strategy and thus validate the effectiveness of the prediction model. We hypothesize that having a fairly correct "oracle" in the form of the prediction model should allow for better trading if it is used effectively. The key would be to find a trading strategy robust to the occasional short comings of the prediction model which could otherwise prove costly and mitigate the positive benefits.

Chapter 7

Conclusion

We experimented with several approaches to predict the monetary success of movies in the box office. Specifically, we attempted to predict movie-stock performance in the HSX prediction market as well as actual box office revenue. This thesis focussed on building models using variables that reflected the input of the collective whole with the addition of some sentiment and social network analyses. The data was drawn from HSX, IMDb, Rotten Tomatoes, and Box Office Mojo.

Initial models included attempts to the direction and then magnitude of daily changes in the HSX movie stock price based on the past history of the movie stock. These were built using linear regression and consensus voting. The results here showed the roots of promise, but we then chose to focus on predicting final performance instead of day-to-day fluctuations.

We then built a naive Bayesian classifier to identify movies as flops, so-so, and blockbusters based on their revenue to production budget ratios. The simple normal-distribution-based classifier used a net sentiment and betweenness score and was more effective than random guessing. While it had some errors, it was rare for it to completely misclassify a movie, i.e. identify a flop as a blockbuster or vice versa.

The most effective models were the single-variable and multi-variable linear regression models built to predict both the HSX delist price and the final box-office gross revenue. We found that the HSX price even 2 weeks before release was a strong proxy for the eventual HSX delist price. Using multilinear regression, we were able to

refine the model to produce slightly more accurate predictions nearly a month before the HSX movie stock was delisted. The multilinear regression beat the best single-variable model by around 10, or 15.18%, and the HSX price itself by 5, or 8.66%, in terms of mean squared error.

A similar approach to predict the final gross revenue also produced fairly accurate productions, with the lowest mean squared error around 213. In this case, predictions were being made over a month before the movie closed in the box office. That error was also skewed by a few severe mispredictions rather than a general model tendency to greatly over or under predict. Here the improvement of the multilinear regression over the best single-variable regression model was much more significant, reaching an improvement of around 100, or 32.00%, by 6 days after release. Furthermore, the improvement over the HSX price exceeded 150, or 43.72%. This second set of results not only shows a greater improvement of our model over HSX, but also shows us two things:

1. Movies seem to continue to gross revenue at the same rate relative to each other before and after four week post-release.

2. We can refine the HSX prediction market price targeted to predicting four-week performance to extrapolate and accurately predict final box-office performance.

Bibliography

[1] Ilan Alon, Min Qi, and Robert J. Sadowski. Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, 8:147–156, 2001.

[2] Joyce Berg, Robert Forsythe, Forrest Nelson, and Thomas Rietz. Results from a dozen years of election futures markets research. In Charles Plott and Vernon Smith, editors, *Handbook of Experimental Economic Results*. Elsvier, Amsterdam, 2001.

[3] Eric Bonabeau. Decisions 2.0: The power of collective intelligence. *MIT Sloan Management Review*, January 2009.

[4] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[5] Sergey Brin and Larry Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, 1998. Elsevier.

[6] P. Gloor, J. Krauss, S. Nann, K. Fischbach, and D. Schoder. Web science 2.0: Identifying trends through semantic social network analysis. In *IEEE Conference on Social Computing*, Vancouver, August 2009.

[7] J. Krauss, S. Nann, D. Simon, K. Fischbach, and P. Gloor. Predicting movie success and academy awards through sentiment and social network analysis. In *Proceedings of European Conference of Information Systems*, Galway, Ireland, June 2008.

[8] B. R. Litman and H. Ahn. Predicting financial success of motion pictures. In B. R. Litman, editor, *The motion picture mega-industry*. Allyn & Bacon Publishing, Inc., Boston, MA, 1998.

[9] Thomas Malone. What is collective intelligence?, October 2006. http://www.socialtext.net/mit-cci-hci/index.cgi?what_is_collective_intelligence.

[10] Hirofumi Matsuzawa and Takeshi Fukuda. Mining structured association patterns from databases. In *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 233–244, London, UK, 2000. Springer-Verlag.

[11] T. Nasukawa, M. Morohashi, and T. Nagano. Customer claim mining: Discovering knowledge in vast amounts of textual data. Technical report, IBM Research, Japan, 1999.

[12] B. Pang and L. Lee. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, July 2002. Associate for Computational Linguistics.

[13] Eldar Sadikov, Aditya Parameswaran, and Petros Venetis. Blogs as predictors of movie success. Technical report, Stanford University, 2009.

[14] M. S. Sawhney and J. Eliashberg. A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, 15(2):113–131, 1996.

[15] Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.*, 30(2):243–254, 2006.

[16] C. J. van Rijsbergen, S. E. Robertson, and M. F. Porter. New models in probabilistic information retrieval. In *British Library Research and Development Report*, number 5587. British Library, London, 1980.

[17] S. Wassermann and K. Faust. *Social Network Analysis.* Cambridge University Press, 1994.

[18] J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

[19] Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:301–304, 2009.