University of Dublin

TRINITY COLLEGE

Sentiment Analysis in Financial Markets

Jonathan Hennessy-Doyle
B.A.(Mod.) Computer Science
Final Year Project April 2015
Supervisor: Prof. Khushid Ahmad

School of Computer Science and Statistics

O'Reilly Institute, Trinity College, Dublin 2, Ireland

Declaration of Authorship

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university.

_____ April 21, 2015

Permission to Lend

I agree that the Library and other agents of the College may lend or copy this thesis upon request.

_____ April 21, 2015

1 Abstract

Machine learning techniques were used to establish a possible relationship between a financial market and the sentiment of news items related to that market. The specific financial market examined was the S&P 500, with data gathered from Yahoo Finance and news sentiment from a wide variety of RSS feeds over a limited time period. Genetic algorithms were used in an attempt to identify a model that mapped sentiment data to market data. The models produced by the genetic algorithm system were evaluated and the best produced model was found to be a reasonably good predictor of the sign of daily returns but a poor predictor of the magnitude of the actual returns. The genetic algorithm system was found to perform reasonably well at producing models to predict prices directly but performed poorly at producing models to accurately predict overall returns. Limitations of the project design were identified and topics for future research highlighted.

2 Acknowledgements

I would like to thank my mother Anne and my father Owen for all of the help, and support they have aided me with during this project. I would also like to thank my project supervisor Prof. Khurshid Ahmad, for his support and suggestions throughout this project.

Contents

List of Figures

List of Tables

3 Introduction

Sentiment can be defined as a thought, view or attitude based primarily on emotion rather than reason (1). A financial market is a physical or virtual place were people and businesses/organisations can trade financial securities, commodities, and other items of value. Usually at relatively low transaction costs and at prices that should reflect the supply and demand of these items (2). As early as the 1930's starting with J.M. Keynes, numerous authors since then have considered the possibility that a significant presence of sentiment-driven investors can cause prices to depart from fundamental financial values (3). The price of these items should be a function of their relative supply and demand, however, given that the financial market is where people and entities are present, and the latter are composed of people then sentiment may influence their decisions. While the majority of traders in the market are considered "rational" those who do not have access to detailed information (some who have been called "outsiders") may be influenced more by sentiment (4). Other authors (5) have identified evidence that investor sentiment plays a significant role in international market volatility.

With significant advances in information technology, that provide instant access and sharing of information, investors can rapidly obtain more valuable and timely information (6). These authors concluded from their study that information of firm-specific news articles can enrich the knowledge of investors and affect their trading activities. Predicting stock market behaviour has always had a certain appeal for researchers, however, numerous attempts have been made but the difficulty has been the inability to model the behaviours of human traders (7).

Behavioural economics tells us that emotions can profoundly affect individual behaviour and decision-making (8) and that in modern times Twitter

sentiment might be used to predict movements of a stock or sector may yield promising insights into potential practical applications (9). These authors investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. They found that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions and furthermore they stated an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error (MAPE) by more than 6%. Other authors (10) have used data from an automated news analytics tool to identify responses in trading returns due to news arrivals. They concluded that sentiment indicators can predict future price trends. Cohen-Charash et. all, (11), examined whether the impact of press reports could predict changes in stock market prices. They found that pleasant moods predicted increases in NASDAQ prices, while activated unpleasant mood predicted decreases in NASDAQ prices.

This project investigated the effectiveness of machine learning as applied to sentiment analysis and its possible relationship to financial markets. The key question for this work was: Can a market be accurately predicted solely by sentiment about that market? The specific objectives of this project were:

1. To build, test and evaluate models that map sentiment from news articles about a market to that market's data.

2. To evaluate the performance of the best fit model.

4 Design

The design of this project involved gathering news articles and subjecting them to analysis using RockSteady. This program then extracted sentiment data

which was then fed into the Genetic Algorithm. Also, inputted into the Genetic Algorithm was financial data which was gathered from Yahoo Finance. The Genetic Algorithm then produced refined models which map sentiment data to financial data (Figure 1).

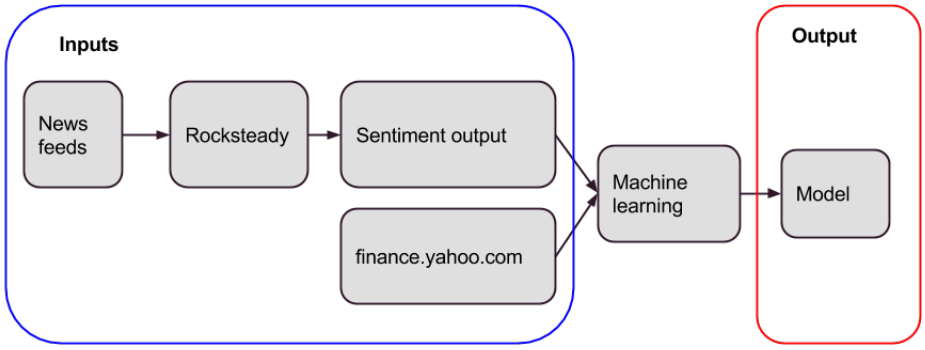

Figure 1: Overview of Project Architecture

This project used a Genetic Algorithm System to build, test and refine models to map sentiment data to financial market data. A genetic algorithm is a machine learning technique that mimics natural evolution to build, test and refine models to fit a specified purpose or purposes. A genetic algorithm system was chosen over regression-equations based techniques do to its resilience against multiple, possibly related, variables mapped to a non linear result. A Genetic algorithm is a search heuristic (12) that mimics the process of natural selection. Genetic algorithms are generally constructed using four main components, an initial population seeding, a fitness-for-purpose test, a selective reproduction system and a generation loop. Genetic Algorithms are useful for generating solutions to optimisation and search problems without a requiring in depth knowledge of the problem at hand.

4.1 Financial Data

The financial data targeted by this system must be in the form of a continuous time series that corresponds to a well defined period of sentiment data. Any market can be used provided that there exists sufficient related sentiment data and that a continuous end of day time series is available.

4.2 Sentiment Data

The sentiment data for this project was produced by the Rocksteady program. Rocksteady takes a large number of news articles and extracts a selected number of sentiment measures from them. This sentiment data is then inputted into the genetic algorithm system. This project required a large number of news articles related to the target market that cover a continuous time period. The quality of the sentiment data increases with more available news articles. News articles published on non-trading days are considered as having been published at 12.00am on the next trading day. The Rocksteady program requires the Java platform to operate. The source news articles can be provided either in LexisNexis text format or as plain text files stored on a local drive or taken from RSS feeds. Financial data can be taken from from any financial service provider, for example finance.yahoo.com. The end of day price is available on this service within a specified date range in CSV format.

4.3 Genetic Algorithm System

The Genetic Algorithm System (Fig. 1) requires four functions which are, an Initial Population Seeding, a Fitness-for-Purpose test, a Next Generation step and a Generation Loop. An initial population, of size Pop was chosen based on time available, is seeded from random values or a known reasonable value. The fitness-for-purpose test provides a quantifiable and comparable measure-

ment of the model. The requirement of this function is that the value of Fitness (A)>Fitness (B) when model B is a superior model than model A, for the tested purpose. Multiple fitness functions for different purposes may be used in a single system provided that there exists a meaningful measurement of the combined fitness values. The next generation function performs natural selection to choose the fittest individuals in the population which are then subjected to mutation and cross breeding to produce the next generation. The percentage of the previous population to target for reproduction, the chance and method of mutation and the chance and method of cross mutation are application defined. The generation loop continues iterating through generations until an application defined "Good Enough" is reached or no more time is available for the computation.

4.3.1 Genetic Algorithm System Example

For example, consider a system which can be represented as a string of 1s and 0s with length 11. A genetic algorithm to find an optimal state of the system could be constructed as follows:

- An initial seeded population of 100 individual states with randomly assigned values, e.g. "100101110000"
- The fitness-for-purpose function for this example is defined assuming that the global optimal solution is "11100000111" and the fitness value is the sum of the differences between a state's and the optimal solution's individual digits. By this definition the value of Fitness ("100101110000") is 10.
- The next generation function is set to select the fittest 10% of the population of each generation and subject them to mutation and cross breeding. The probability of mutation and cross breeding with each generation is set to 0.5.

- The mutation method is defined as : Select a random digit from the state and invert it.
- The cross breeding method is defined as follows: Select a random neighbouring pair of digits from the state. Then replace the selected pair with the corresponding pair from a randomly selected other state.
- The Generation Loop is defined as follows: "Good Enough" is defined as Fitness(state) = 1. This represents a single digit difference compared with the optimal solution. The maximum generations to iterate through is defined as 10

5 Implementation

The methods and data used during the project are detailed and a description of the implementation of the project design is described in this section. A number of challenges were encountered and the implementation is discussed and contrasted with possible alternative approaches.

5.1 Sentiment Data

The Sentiment Data was produced by the Rocksteady program. Rocksteady is a Java based application that takes news articles in several formats and extracts ten sentiment measures. In this project these included the number of: articles, terms, positive terms, negative terms, active terms, passive terms, strong terms, weak terms, and finally economic, political and military terms. These ten sentiment measures were combined by day and outputted to a CSV file. This CSV file was used as input to BEBOP.PY. Rocksteady was configured to download news articles from RSS feeds. RSS feeds have limited available back catalogue with articles often being retrievable for 24 hours or less. To address this the

articles taken from the RSS feeds were stored in a local corpus until needed. See Appendix A which contains further data on the news articles used in this project.

5.2 Financial Data

The financial data targeted in this project were the end of day prices and end of day returns taken from Yahoo's finance site's listings for the S&P 500 (B). The S&P 500 was chosen due to the ready availability of financial data and the large number of articles published concerning the membership of the market. This data is available for download as a CSV file the relevant time series for the market within a selected range, for this project the range was 01.03.2015 to 17.04.2015. The end of day price and end of day returns were chosen as the targets for the models as they allow the analysis to consider the trading and articles published that day to be considered as a single unit of discrete time. The end of day price and end of day returns are closely related and are targeted only separately due to differing performance of the genetic algorithm system in producing models for each.

5.3 Genetic Algorithm System

The genetic algorithm system outlined above was implemented in a single Python file, BEBOP.PY. This file reads in the sentiment data CSV file and the financial data CSV file described above and outputs the best fit model found by that iteration of the genetic algorithm system. The time series generated by the best fit model is displayed, along with actual market data, using mathplotlib[1]. The parameters chosen for the genetic algorithm system are detailed below:

- A model was defined as a list of 10 coefficients corresponding to each of the

[1] http://matplotlib.org/

sentiment measures outlined above. The coefficients represent the relative weight of each measure in predicting the target value.

- The initial population was selected to contain 1000 randomly generated models.
- The fitness for purpose test was defined as the sum of the square of the difference between the predicted values and the actual values.
- The next generation function was defined as selecting the fittest 10% of the previous generation for reproduction. Each model so selected was used for reproduction roughly ten times so the size of each generation remains constant.
- The mutation rate was defined as 0.5 with the mutation method being to select a random coefficient from the model and to replace it was a random value.
- The cross breeding rate was defined as 0.5 and the cross breeding method was defined as replace a random pair of coefficients from this model with the corresponding coefficients from another model selected for reproduction.
- The generation loop was defined to terminate after 10 generations. This is sufficient time for the system to reach a state where any further improvements will require exponential cost. The "Good Enough" target was defined as the equivalent of an average daily error of 0.5%. No model reached this target.

6 Results

To demonstrate the results of this project the program BEBOP.PY was run on sentiment data (Appendix 1) and financial data for 35 trading days from 01.03.2015 to 17.04.2015. The run time for this system was 56 seconds. Model A[2] which was output as the best generated model from a run of genetic algorithm system predicted the the end of day prices with an average error of 4.33% (Fig. 2). In contrast Model B[3], produced from a different run of BEBOP.PY, was more successful in predicting the sign of end of day returns (Fig. 3).

[2]Model A = [-7.0915005351209315, -5.8716038035739615, -17.544869328198264, -2.518193360683571, 2.088560502746449, -17.410354039707524, 6.5514344025523465, -0.15729057896132628, 4.187067772603826, -8.0543931785375]

[3]Model B = [0.011484472805949782, -0.0051601653957655345, 0.0055480592654472 1, -0.0011227052919888373, -0.015575352728237433, 0.0014521124016208184, 0.0009964638009343357, -0.00817366220772595, -0.004639328886936736, -0.014426040782246225]

6.1 End of Day Price Prediction

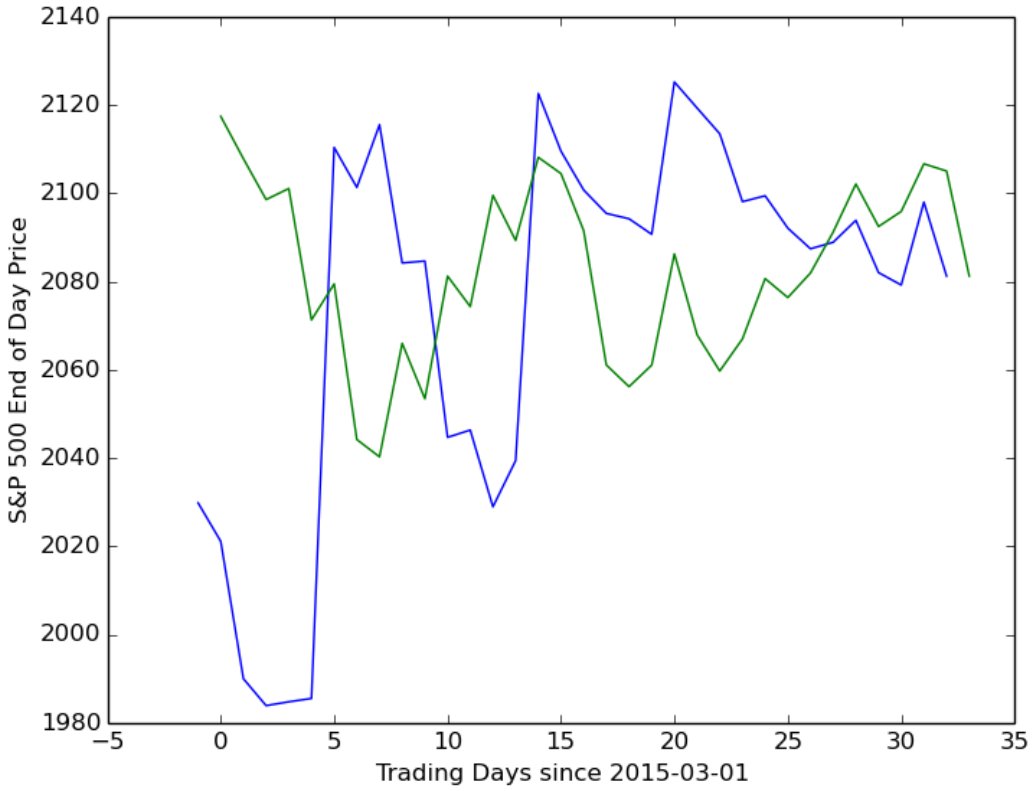

Figure 2: End of Day Price Prediction in the S&P 500 with Sentiment Data from 01.03.2015, based on Model A. The Blue line is the predicted end of day price. Green line is the actual end of day price.

Figure 2 shows several notable occasions, days 2, 6 and 12 for example, where the magnitude of the predicted price differed substantially from the actual end of day price. These differences prevent Model A from being economically viable. However, the sign of the direction of change predicted by Model A was correctly 56%[4] of the time.

[4] 18 correct predictions over 32 day changes.

6.2 End of Day Returns Prediction

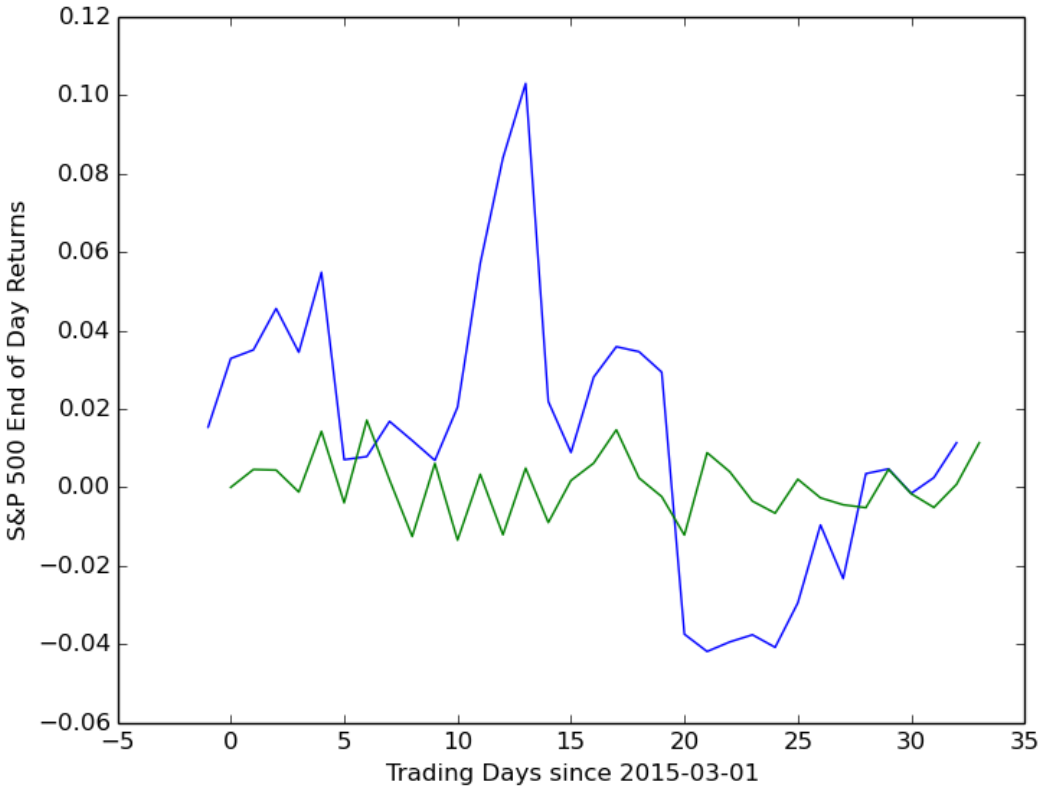

Figure 3: End of Day Returns Prediction in the S&P 500 with Sentiment Data from 01.03.2015, based on Model B. The Blue line is the predicted end of day price. Green line is the actual end of day price.

Figure 3 shows that the magnitude of the return is also often significantly in error, most notably on day 14. However, the sign of the return, representing the direction of change in the market compared to the previous day, was predicted with 63%[5] accuracy, significantly more than would be expected from random variation.

[5] 20 correct predictions over 32 day changes.

7 Evaluation

The results indicated that the sign of the predicted returns were somewhat well predicted, whereas the magnitude of any change in market prices was not. The results for end of day returns prediction indicated that the generated model depicted therein can make predictions on the sign of the returns that are significantly better than what would be expected from a random coin flip. However, the results further indicated that there were several days in which the magnitude of the predicted return was inaccurate, and so the predicted prices are significantly different from the actual market values. This makes the model unsuitable for real world predictions.

7.1 Genetic Algorithm System

The performance of the genetic algorithm system was not as efficient as was expected. There was an unexpectedly low generation on generation improvement compared to typical genetic algorithm systems. This issue is possibly due to the parameters used when mutating a state not considering the previous state's corresponding value, and thereby losing progress, although such an effect has not been previously noted.

7.2 Sentiment Data

The quality of the sentiment data suffered from a lack of news articles before March 2015 due to the issue relating to a 24 hour limit on the availability of RSS feeds. This is a major limiting factor as models can only be evaluated over the 34 trading days with sufficient corresponding sentiment data. Regarding the financial data there were no issues with acquiring or incorporating the financial data into the system used.

8 Conclusions

This project investigated the relationship between a market (S&P 500) and the sentiment of news around that market in a given time period. This project used genetic algorithms to attempt to identify a model that mapped sentiment data to market data. The models produced by the genetic algorithm system were evaluated and the best produced model was found to be a reasonably good predictor of the sign of daily returns but a poor predictor of the magnitude of the returns. The genetic algorithm system was found to perform reasonably well at producing models to predict prices directly but performed poorly at producing models to accurately predict returns directly. The best found model had an average daily error of 2.36% when predicting end of day returns and 4.33% when predicting end of day prices. The best found model was not economically viable. The project faced a number of significant challenges in particular a shortage of sentiment data with which to train and evaluate the models. However, the results confirm that the methods used in this project were found to be promising where they relate to the prediction of the sign of financial returns. Future work in this area would benefit from a larger pool of data, particularly sentiment data, with which to train and test the models. The system may also be modified to predict returns directly, ideally with similar or superior results to predicted prices.

References

[1] *Sentiments A definition.* thefreedictionary.com, 2015.

[2] M. Groz, *Forbes Guide to the Markets.* John Wiley & Sons, Inc., 2009.

[3] Y. J. Stambaugh, R.F. and Y. Yuan, *The Short of It: Investor Sentiment and Anomalies.* Journal of Financial Economics, doi:10.1016/j.jfineco.2011.12.001, 2011.

[4] B. Mendel and A. Shleifer, *Chasing noise.* Elsevier, 2012.

[5] J. Baker, P. Wurgler and Y. Yuan, *Global, Local and Contagious Investor Sentiment.* Journal of Financial Economics. Vol 104 00 272 to 287, 2012.

[6] P. L. L. L. Q. G. Y. C. Q. Li, T. Wang, *The Effect of News and Public Mood on Stock Movements.* Inform. Sci., 278, pp. 826–840, 2014.

[7] Z. Y. H. C. Schumaker, R.P. and C. H, *Evaluating Sentiment in Financial News Articles.* Decision Support Systems, Volume 53 Issue 3, June, 2012, p. 458-464, 2012.

[8] J. H. M. H. Bollen and X. Z. X, *Twitter Mood Predicts the Stock Market.* Journal of Computational Science 2 (2011) 1–8. doi:10.1016/j.jocs.2010.12.007, 2011.

[9] E. D. Brown, *Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market.* SAIS 2012 Proceedings. Paper 7, 2012.

[10] A. Groß-Klußmann and N. Hautsch., *When Machines Read the News: Using Automated Text analytics to Quantify High Frequency News-implied Market Reactions.* Journal of Empirical Finance 18:321–40, 2010.

[11] K.-M. J. S. B. GCohen-Charash Y, Scherbaum CA, *Mood and the Market: Can Press Reports of Investors' Mood Predict Stock Prices?* PLoS ONE 8(8): e72031. doi:10.1371/journal.pone.0072031, 2013.

[12] M. Mitchell, *An Introduction to Genetic Algorithms*. MA: MIT Press, 1996.

A Sentiment Data

The sentiment data used in this project is too large to display here, and instead can be found in the file "sentiment-2015-4-17.out" on the CD accompanying this article. The data is stored in the form "Title,Date of First Article, #Articles, Terms, Positiv, Active, Passive, Strong, Weak, Econ@, POLIT, Milit" with each row representing a day on which an article, or articles, were published and the values being the results of sentiment analysis on the day's combined corpus. This data was produced by RockSteady from 134 RSS feeds which contained a total of 9962 articles published between 01-03-15 and 17-04-15.

B Financial Data

Table 1: S&P500 from 01-3-15 to 17-04-15

Date	Open	High	Low	Close	Volume	Adj Close
2015-04-17	2102.58008	2102.58008	2072.37012	2081.17993	3627600000	2081.17993
2015-04-16	2105.95996	2111.30005	2100.02002	2104.98999	3434120000	2104.98999
2015-04-15	2097.82007	2111.90991	2097.82007	2106.62988	4013760000	2106.62988
2015-04-14	2092.28003	2098.62012	2083.23999	2095.84009	3301270000	2095.84009
2015-04-13	2102.03003	2107.6499	2092.33008	2092.42993	2908420000	2092.42993
2015-04-10	2091.51001	2102.61011	2091.51001	2102.06006	536200000	2102.06006
2015-04-09	2081.29004	2093.31006	2074.29004	2091.17993	3172360000	2091.17993
2015-04-08	2076.93994	2086.68994	2073.30005	2081.8999	3265330000	2081.8999
2015-04-07	2080.79004	2089.81006	2076.1001	2076.33008	3065510000	2076.33008
2015-04-06	2064.87012	2086.98999	2056.52002	2080.62012	3302970000	2080.62012
2015-04-02	2060.03003	2072.16992	2057.32007	2066.95996	3095960000	2066.95996
2015-04-01	2067.62988	2067.62988	2048.37988	2059.68994	3543270000	2059.68994
2015-03-31	2084.05005	2084.05005	2067.04004	2067.88989	3376550000	2067.88989
2015-03-30	2064.11011	2088.96997	2064.11011	2086.23999	2917690000	2086.23999
2015-03-27	2055.78003	2062.83008	2052.95996	2061.02002	3008550000	2061.02002
2015-03-26	2059.93994	2067.1499	2045.50	2056.1499	3510670000	2056.1499
2015-03-25	2093.1001	2097.42993	2061.05005	2061.05005	3521140000	2061.05005
2015-03-24	2103.93994	2107.62988	2091.50	2091.50	3189820000	2091.50
2015-03-23	2107.98999	2114.86011	2104.41992	2104.41992	3267960000	2104.41992
2015-03-20	2090.32007	2113.91992	2090.32007	2108.1001	5554120000	2108.1001
2015-03-19	2098.68994	2098.68994	2085.56006	2089.27002	3305220000	2089.27002
2015-03-18	2072.84009	2106.8501	2061.22998	2099.50	4128210000	2099.50
2015-03-17	2080.59009	2080.59009	2065.08008	2074.28003	3221840000	2074.28003
2015-03-16	2055.3501	2081.40991	2055.3501	2081.18994	3295600000	2081.18994
2015-03-13	2064.56006	2064.56006	2041.17004	2053.3999	3498560000	2053.3999
2015-03-12	2041.09998	2066.40991	2041.09998	2065.94995	3405860000	2065.94995
2015-03-11	2044.68994	2050.08008	2039.68994	2040.23999	3406570000	2040.23999
2015-03-10	2076.13989	2076.13989	2044.16003	2044.16003	3668900000	2044.16003
2015-03-09	2072.25	2083.48999	2072.20996	2079.42993	3349090000	2079.42993
2015-03-06	2100.90991	2100.90991	2067.27002	2071.26001	3853570000	2071.26001
2015-03-05	2098.54004	2104.25	2095.21997	2101.04004	3103030000	2101.04004
2015-03-04	2107.71997	2107.71997	2094.48999	2098.53003	3421110000	2098.53003
2015-03-03	2115.76001	2115.76001	2098.26001	2107.78003	3262300000	2107.78003

C CD Contents

The accompanying cd contains three files, "bebopy.py", "sentiment-2015-4-17.out", and "s&p500-mar-1-to-apr-17.csv". bebop.py contains the executable program which may be invoked by "python bebop.py". This will produce a graph similar to Fig. 2 and the console output will contain the fittest model found in each generation. "sentiment-2015-4-17.out" contains the sentiment data in CSV format as described in Appendix A. "s&p500-mar-1-to-apr-17.csv" contains the financial data shown in Appendix B in CSV format.