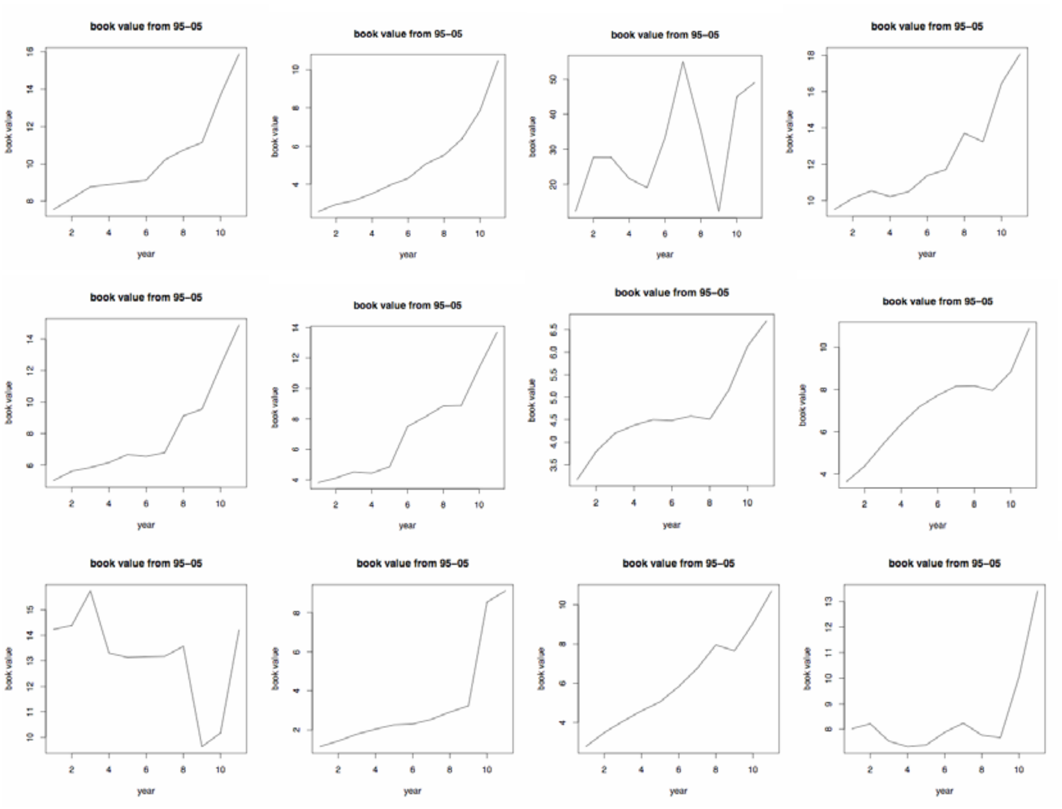
MKV/BKV Ratio Prediction

Yan Sui Zheng Chai

Introduction:

One of the major applications of data mining is in the financial sector, specifically, the stock market. Accurate predictions of the stock market prices, though seems appealing, is an extremely difficult task. Instead of tackling this challenge directly, we try to solve a related prediction problem, namely, predicting the ratio of market value over book value.

Market value (MKV) is the stock price driven by the market, whereas the book value (BKV) is the accounting value of a company's stock. This ratio indicates the investors' confidence in the company, or the investors' expectations of the company's performance. The ratio greater than one means that the investors are confident about a company's potential. On the other hand, the ratio less that one indicates that investors expect this company to do worse in the near future. Since the accounting value of a company is updated annually and generally stays relatively constant throughout the year, accurate prediction of the MKV/BKV ratio provides rough estimate of the MKV. The following graphs plot the MKV change from 1995 to 2004 for a subset of Dow 30 companies.



While BKV is relatively predictable, the MKV is influenced by close to infinite number of factors. Ideally, all of the factors should be taken into account in predicting the MKV or MKV/BKV. Obviously, this is not feasible in practice. Feature selection must be used to reduce the number of dimensions of input data. The purpose of feature selection is to select the most influential factors and ignore the less influential ones, in order to reduce the complexity of prediction algorithms. Once the most important factors are selected, we can then proceed to do the prediction.

We first train a predictor using historical data and then use this predictor to predict the ratio for the given data of a more recent year. The accuracy of the prediction could be measured by comparing the predicted ratio with the actual ratio of the year.

Feature Selection:

Two main feature selection algorithms are considered: genetic algorithm and greedy algorithm.

The genetic algorithm is search and optimization techniques based on Darwin's principle of Natural Selection. Let's first define some terms. An individual is a set of attributes. Each iteration of the algorithm is called a generation. Fitness value is a measure of how good of solution an individual is. A gene is an attribute of a solution. Initially, the algorithm randomly selects a number of individuals. The fitness value is calculated for each individual. After each generation, individuals with worse fitness value die off and new generation of individuals caring the genes of one or more remaining individuals are introduced to the population. This process repeats itself until the max number of generations is reached or the satisfactory solution is found. The adaptive behavior of genetic algorithm ensures the convergence of the solution without exhaustively search. The major weakness of the algorithm is that it does not consider the orthogonality of chosen attributes. The attributes chosen could have significant overlap.

The greedy algorithm tries to pick attributes that are orthogonal to one another. It is also an iterative procedure. It keeps a list of chosen attributes. At each iteration, the algorithm picks the attribute that adds the most power to the existing solution. Therefore, the chosen attributes are as orthogonal to one another as possible. The greedy algorithm tries to find local optimal solutions. If two attributes together have great predictive power, but do not when considered individually, the two attributes are likely left of the final solution. In addition, the time complexity of greedy algorithm is n^2.

Data Used:

The experiment is focused on the Dow Jones Industrial Average (Dow 30) stocks, which consists 30 of the largest and most widely held public companies in the United States. Attributes are from CRSP/COMPUSTAT merged database and MKV/BKV ratio is from COMPUSTAT North America database of Wharton's wrds.

For each company, one MKV/BKV ratio is collected per year. The high price, low price, and the last day closing price of each month are averaged to become the estimated MKV for the year.

The most recent change in the Dow Jones Industrial Average occurred on November 1, 1999. Here is the current list of the Dow Jones Industrial Average:

AT&T
Alcoa
American Express
Boeing
Caterpillar
Citigroup
Coca Cola
Disney
Du Pont
Eastman Kodak
Exxon
General Electric
General Motors
Hewlett Packard
Home Depot
Honeywell International Inc.
Intel Corporation
International Business Machines
International Paper
Johnson & Johnson
McDonald's Corporation
Merck & Company
Microsoft Corporation
Minnesota Mining & Manufacturing
Morgan J. P.
Philip Morris Companies
Procter & Gamble
SBC Communications
United Technologies
Wal-Mart Stores

Performance Results:

Since both feature selection algorithms have their own advantages and disadvantages, we built predictors using both algorithms. The performance on the Dow 30 stocks is shown in the following figures.

The input to feature selection is N, the number of features to use. A Support Vector Machine (SVM) is trained for each stock, using selected feature data for the firm. Training data used are from 1995 to 2004. This SVM is then used to predict the MKV/BKV ratio for the year 2005. The relative prediction error of the prediction is calculated as follows.

$$\frac{|\,predicted\_value - target\_value\,|}{target\_value}$$

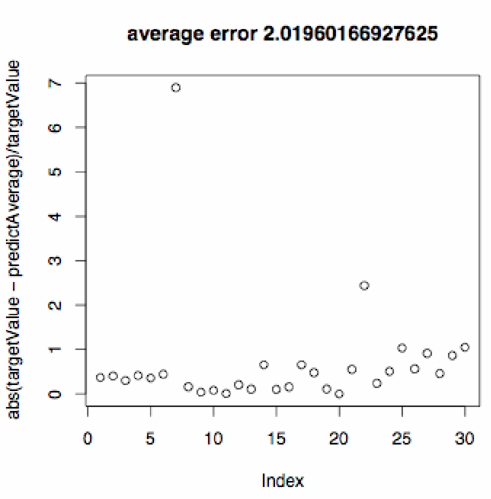
If the predicted value equals target value, the error will be zero. On the other hand, if the predicted value is very different from the target value, the error will be a positive number.

Note: Our assumption is that one stock's performance on the market influences other stocks' performance in the same year. Therefore, we build one SVM for all 30 stocks, instead of considering all stocks independently.

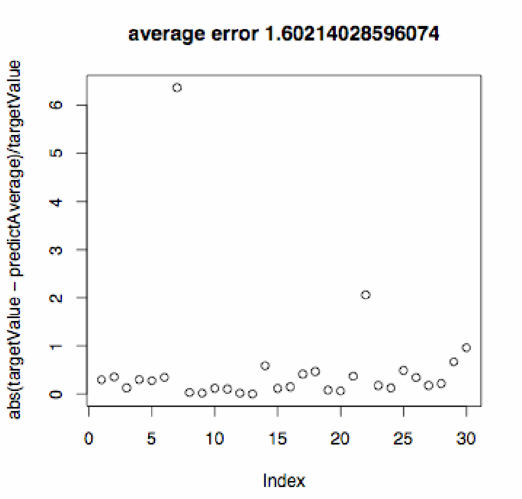
Genetic Algorithm:

First, let's look at the performance using Genetic Algorithm. For our following experiments, we used N = (5, 10, 20, 30) and the results are as shown in the following figures.

N = 10

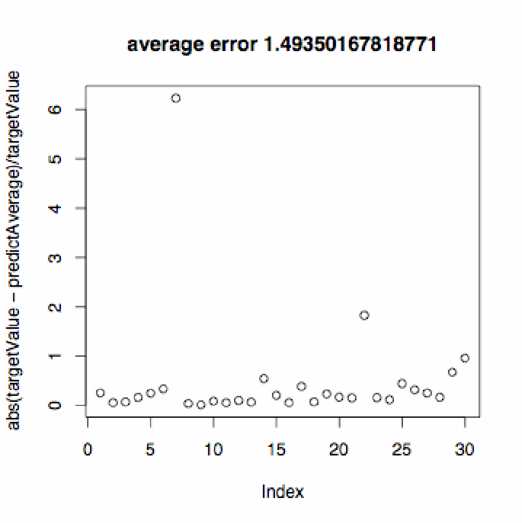


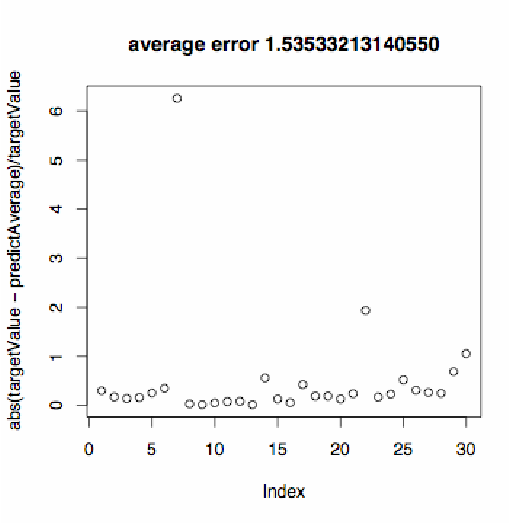
N = 20



N = 30

N = 40



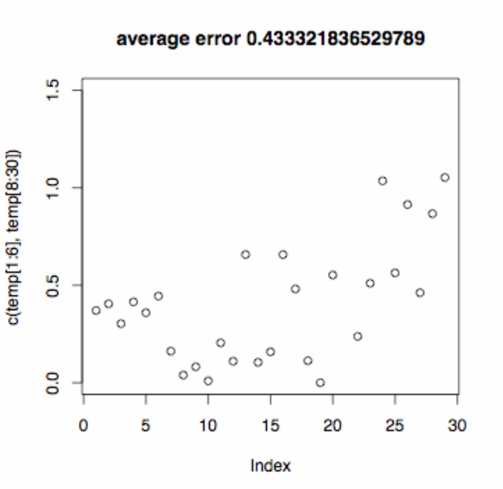


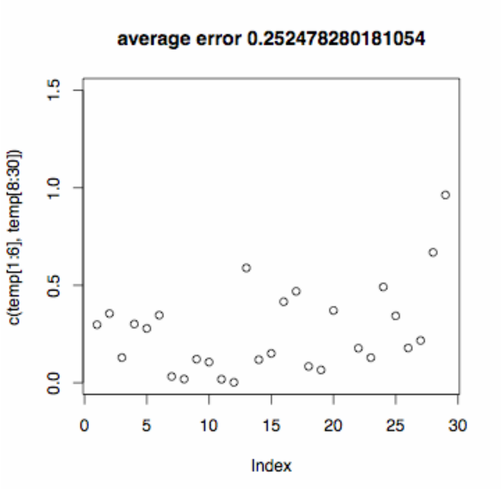
Because of the GA's randomness, the error is slightly different after each run. Therefore, the above figures are the average of 10 runs for each value of N. Also, in the above figures, there is an outlier, namely General Motors. Our SVM significantly over-estimated this stock. As the result, the error is up around 6 for GM, whereas none others have error greater than 2.

If we take out this outlier, the result looks like this.

N = 10

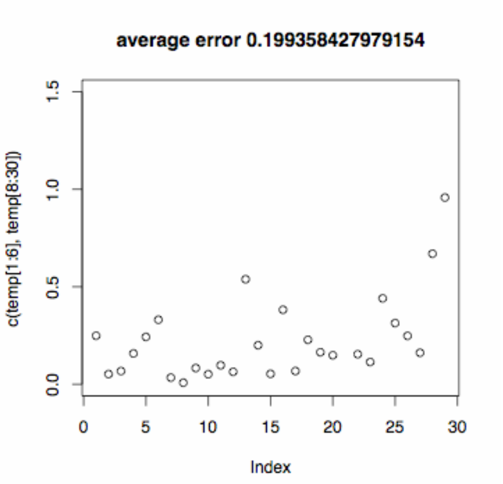
N = 20

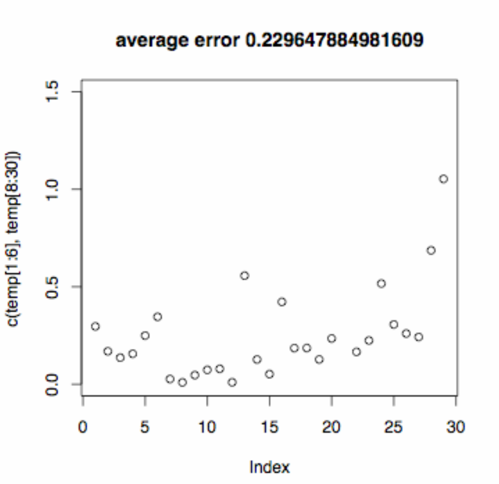




N = 30

N = 40





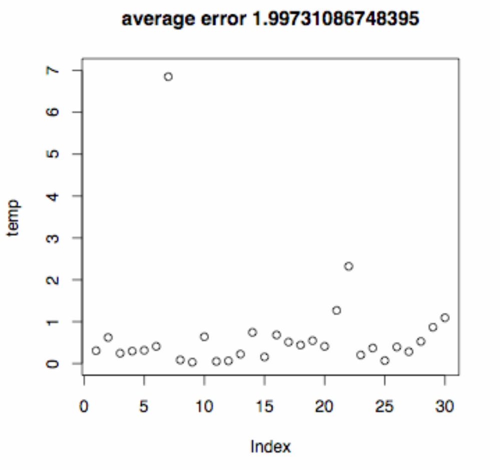
The SVM seems to perform better with larger N, until N reaches 40. This is due to the fact that more insignificant features are included in the model and over-fitting occurs.

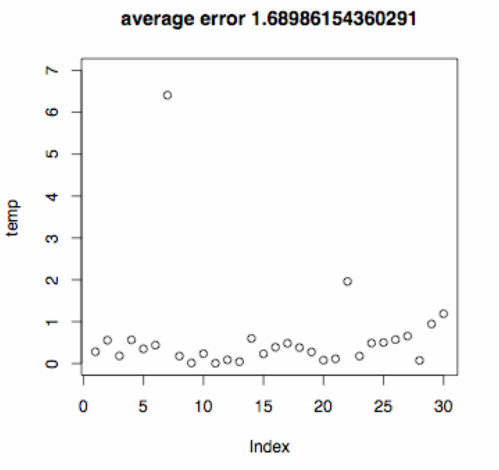
Greedy Algorithm:

We did the same experiment using Greedy Algorithm. The results are shown as follows.

N = 10

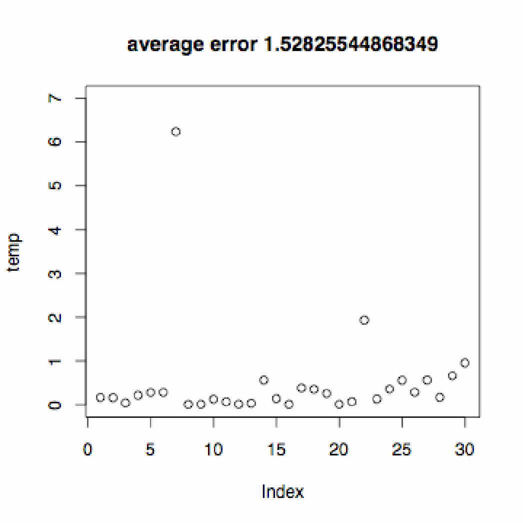
N = 20

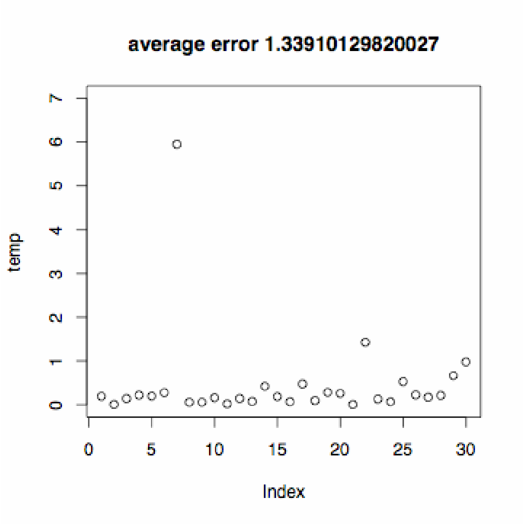




N = 30

N = 40

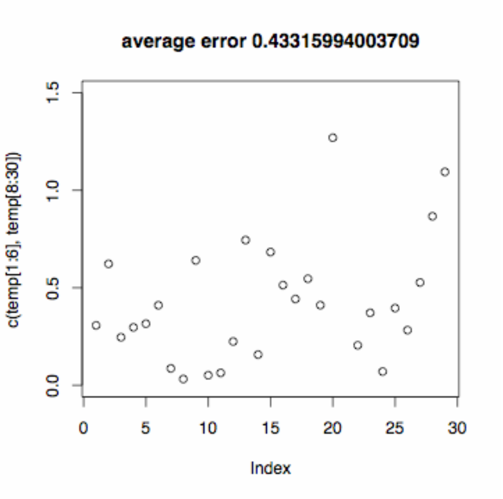


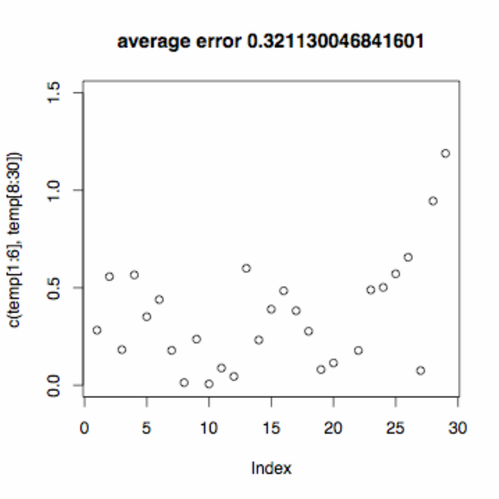


As it appeared in the genetic algorithm, General Motors is the outlier. After taking it out, the figures look like the following.

N = 10

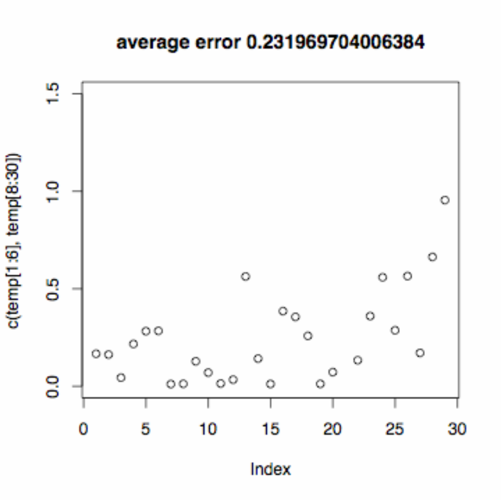
N = 20

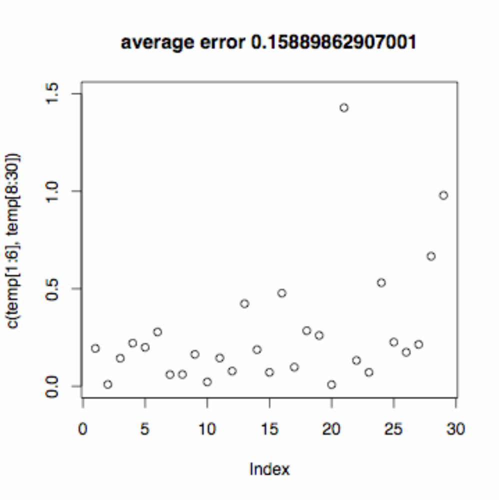




N = 30

N = 40





When using Greedy Algorithm, the performance becomes better as N increases. This is expected since the algorithm only picks additional attributes to increase its accuracy. But as N grows, the contribution of the newly added attributes to the accuracy decreases, as shown in the above figures.

Both algorithms do a poor job on General Motors. This is expected because our model is built for all 30 stocks. Out of the 30 stocks, GM is the only one with ratio < 1. Since the majority of the stocks are performing well. The model is greatly biased. As the result, the prediction for the outlier company, GM, is off by a significant amount.

Conclusion:

Indeed, the task of predicting stock price is extremely complex and challenging. There are numerous factors, which influences the market price of stocks. It may not be feasible to collect all the information necessary to do an accurate regression. For example, if an important policy change took place and boosted or shaken the investors confidence, our model would not be able to detect such factor.

From stock to stock, the prediction accuracy also changes dramatically. As shown in the previous figures, our model significantly over-estimates the value of General Motors, whereas in reality, the investors believe otherwise.

In order to significantly increase the accuracy of such stock price prediction models, more attribute data is needed. This, by itself is already quite a challenging task.