Predicting Excess Equity Returns Using Company Fundamentals

Jason Victor

June 2, 2009

1 Introduction

Most quantitative approaches to stock market analysis are referred to in the finance industry as "technical" approaches—that is, they rely solely on continuously observable market data in order to make bets. Thus, much of the reasoning that has gone into such approaches has been in the spirit of the *arbitrageur*, detecting subtle mispricing anomalies that can be exploited in the extremely short-term, often with little risk involved.

I am approaching the problem from a different standpoint. As opposed to relying on continuously observable market variables—stock prices, credit spreads, bond yields and so forth—I am focusing on a discrete data set: the financial statements of individual companies. Using this information, I will attempt to predict the return of that company's common stock over the S&P 500 index over the next year. In industry parlance, this value is known as *alpha*.

I am evaluating three machine learning models for this regression task: kernel support vector machines, k-nearest neighbors and decision trees.

2 Features

The best quantitative description of an individual firm's performance and operations is in a set of annual SEC filings known as financial statements. Three documents make up a company's financials: the cashflow statement, the income statement and the balance sheet. Each provides a different look at the financial fundamentals underlying the firm's operations. For example, the balance sheet provides detailed information on the assets and liabilities of the company, while the income statement enumerates the components that factor into the company's profit (or loss). Figure 1 illustrates the cycle of cashflows within an individual firm, from financing activities (raising capital in the form of debt or equity) to investments and, finally, a profit.

Because these are raw values (in dollars), an SVM trained on the financials of any one company would not generalize to another company of a different size. Thus, in order to compare one company to another effectively, it is necessary

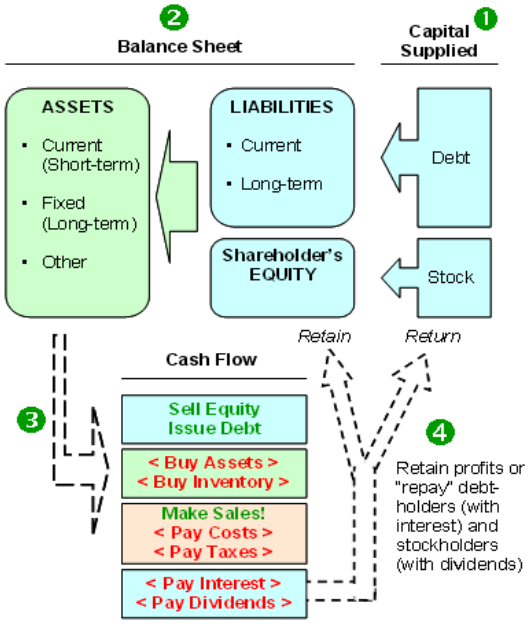

Figure 1: Cashflow of an individual firm

to observe only *ratios* of these figures. These are known as *financial ratios*, and are a common tool in analyzing prospective investments.

I use several types of financial ratios in my feature vector.

- **Liquidity measurement ratios** describe a company's access to cash and ability to cover its short-term obligations
- **Profitability indicator ratios** explain how well the company utilized its resources to generate shareholder value
- **Debt ratios** describe a company's debt load and mix of debt and equity financing
- **Investment valuation ratios** shed light on the price of a company's stock in the open market relative to the performance and breakup value of the firm

Furthermore, I include in the feature vector the excess return (alpha) of the given company's stock over the previous year.

3 Results

To test my method, I used the past three years of information for 50 prominent companies across several sectors. All test companies are industry leaders

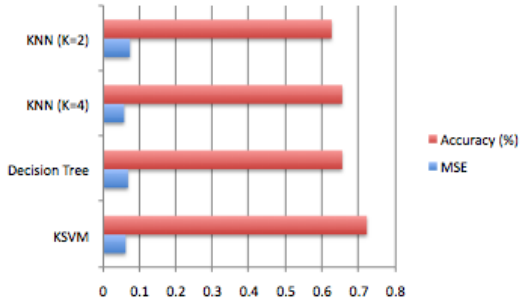

Figure 2: Accuracy and MSE by model type

in market capitalization. To test the performance of a KSVM, I used the implementation available in the Kernlab package for R, with a Gaussian RBF kernel and automatic parameter tuning. I also used the recurisve partitioning routines in the rpart package for R to implement my decision trees. My k-NN implementation is hand-written. A summary of results is available in Figure 2.

I used leave-one-out cross-validation to test the models. For each model, 104 separate tests were performed. The best mean squared error was 0.061, achieved by k-NN for $k = 4$, followed by KSVM with 0.061, decision trees with 0.0675, and finally 0.073 for k-NN with $k = 2$.

However, what is perhaps more interesting from the point of view of portfolio optimization is the *accuracy* of the model. In our case, we define accuracy as the frequency with which the model correctly identified the direction of the output (i.e. whether the company under- or over-performed the S&P 500 index). This objective is more relevant in the investing domain than a simple MSE calculation. The highest accuracy was achieved by KSVM with 72.12%, followed by decision trees and k-NN ($k = 4$) with 65.38%.

I also broke the individual companies into groups based on the main drivers of success. I grouped auto, tech and pharma (for example, Daimler-Chrysler, Google and Novartis) together because innovation combined with marketing is the key to success for these companies. I grouped financial institutions with major integrated oil and gas companies (including Goldman Sachs Group and Exxon-Mobil), because these firms are very much intermediaries, and thus their success is particularly dependent on the quality of their balance sheet. Consumer products (like Proctor & Gamble) and industrials (like Dupont) are on their own. Figure 3 illustrates the accuracy of the KSVM model, broken down by these industry groupings.

The highest accuracy rate was achieved with financials and oil and gas. This makes sense based on domain intuition—since these firms are essentially intermediaries, their ability to make a profit is directly linked to their access to cash. Consumer products came in second; this also makes objective sense, since consumer product manufacturers aren't competing in the domain of innovation so much as that of brand recognition and distribution capabilities. It similarly

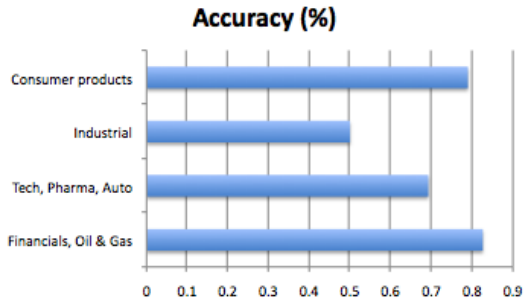

Figure 3: KSVM accuracy by sector

makes sense that auto/tech/pharma would perform poorly, because their success is linked to innovation, which is only loosely correlated to the quantitative indicators described above.

4 Conclusions

Although k-NN was more accurate with respect to a basic MSE calculation, when I framed the problem as a binary classification, KSVM performed far better. Whereas the difference in MSE was negligible, the difference in accuracy was rather broad. Therefore, I conclude that a kernel support vector machine is the choice model for this task.

I am particularly interested in further exploring industry verticals in which this technique works particularly well. As mentioned earlier, financials, oil and gas and consumer products companies allowed a higher degree of accuracy than companies whose businesses are based on more complicated strategy. This suggests that perhaps even industry-specific data could be integrated to build a robust model. Similarly, industries that were difficult to predict could potentially be understood by integrating ratios involving R&D spending into the model.

5 References

Z. Huang, H. Chen, C. Hsu, W. Chen, and S. Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37:543–558, 2004.

X. Hui and J. Sun. An Application of Support Vector Machine to Companies' Financial Distress Prediction. *Modeling Decisions for Artificical Intelligence*, pages 274–282, 2006.

W. Hardle, R. Moro, and D. Shafer. Predicting Bankruptcy with Support Vector Machines. Discussion Paper SFB 649 "Economic Risk", Humboldt University, 2005.