Examining Long-Term Trends in Company Fundamentals Data

Michael Dickens

MOTIVATION

The equities market is generally considered to be efficient, but there are a few indicators that are known to have some predictive power over future price changes. This suggests that the market has some room for identifying inefficiencies. Much work is done on applying machine learning to short-term trading, but there exists little research on using machine learning to detect long-term inefficiencies. Almost all research on using machine learning to make long-term bets about the market. Therefore, we have reasons to believe that there exist long-term market inefficiencies which can be found with machine learning.

FEATURES

To identify useful derived features, I took a set of basic features and computed all possible ratios of these features. Ideally, I would like to compute ratios of sums of features as well as some other more complex combinations at basic features, but this becomes prohibitively expensive as the number of possible derived features expands rapidly.

To test the predictive power of derived features, I used a scoring function that divides the stock market into decades according to a given feature, then returns the absolute value difference in average risk-adjusted returns between the first and last decades. I chose this scoring function because it most accurately illuminates exploitable market inefficiencies.

I was not able to find any features that had stronger predictive power than well-known indicators. Earnings yield performed by far the best, followed by beta and by other ratios that are very similar to the earnings yield. A few other important ratios did have reasonably strong predictive power, but none did as well as well-known features such as earnings yield and return on capital.

LINEAR REGRESSION

Linear regression weekly outperforms both earnings yield and dividend yield, but regression alone does not have better risk-adjusted return. I attempted some modifications to the regression but was not able to produce substantially better results.

Unlike with SVM, I found that adding more indicators as features to the regression did not reduce accuracy. The regression algorithm assigned weight 0 or near-0 to most features. Therefore, the regression algorithm assigns a weight of 0 or near-0 to most features.

SVM

Predicting returns more naturally fits with regression than with classification. However, we want to solve a classification problem in some sense: we want to classify stocks as "buy" or "sell." To do that, we used a threshold (I found the best results when I used 25%) and a negative example otherwise. Then we run the SVM algorithm to separate the set of positive and negative examples.

As with linear regression, I used well-known indicators as features, such as earnings yield, dividend yield, and returns. I also tried using basic fundamentals from companies' annual statements and balance sheets, and then applying kernel methods to produce derived indicators. I found that the SVM did not generalize well to new data, but it did perform well on the training set. Using a small number of well-known features known to have classification that has better predictive power from earnings yield alone.