Research Process

Wenzel Analytics Inc

Introduction

At the end of a process of doing stock selection research for three or four weeks, I wrote this summary of assumptions, detailed research steps and observations from having done the research process and made a presentation to the Twin Cities Chapter of AAII Systematic Investing Special Interest Group on June 15, 2010. The actual data and screens developed are available, but are not prepared such as to make sense as stand-alone documents.

I. Market Assumptions

1. What has worked in the past is likely to work in the future.
2. What has worked in the recent past is more likely to work than what worked in the distant past.
3. There is no reason to expect the stock of companies with "strong" fundamentals to give higher returns than companies with "weak" fundamentals.
    a. Price changes are not driven by any specific company characteristic.
    b. Price is driven strictly by the balance of supply and demand for a stock.
    c. An overvalued quality stock will be a loser, while a grossly undervalued weak stock will show gains.
    d. [The Piotroski scale of nine fundamental characteristics of strong companies is a measurement of company quality, with nine being the highest and one being the lowest quality. The Piotroski screen calls for a score of nine. However, on monthly returns over the last four years, stocks with scores of 1, 8 or 9 did the worst, while stocks in the middle range did significantly better (significance > 000000).]
    e. We are researching fades and the sociology of sentiment.
4. Just willy-nilly exploration doesn't work.
    a. One needs to combine some hypotheses with some data mining exploration.
    b. Cutoff values to use on a variable are mostly discovered through exploration (data mining).
    c. It is easier and more productive to attach an explanation to findings than to start with an explanation and try to find supporting data.

II. Research Steps

A. Prepare Data

1. Collect the data
    a. Download weekly and/or monthly Stock Investor Pro (SIP) data (and the program) from AAII, each to a separate folder.
    b. Create a view consisting of the desired fields for export.
    c. Backup the user settings.
    d. Restore the new user settings to each weekly and/or monthly data set and calculate custom fields.
    e. Export a file from each week and/or month.
    f. I only export rows with Price >= $1, Average Daily 10Day Volume >=5(000), and not equal to Over the Counter.
2. Assemble the data

Lee Wenzel
(952) 944-2699
Lee@WenzelAnalytics.com
www.WenzelAnalytics.com

Wenzel Analytics, Inc.
Registered Investment Advisor
8666 Westwind Circle
Eden Prairie, MN 55344

a. Append each of the files exported from SIP to a single file (table), adding a column for the number of the week or month. (A sequential number is easier to manage than a date.)
b. I had four years of monthly data which came to about 230,000 rows, large enough such that it takes some time for Excel to open files, save files and do calculations (at least on my machine).
c. Create a separate worksheet for dependent variables (returns) so that you can link the data row with consequent returns.
   1). For weekly returns, I add a column to the aggregated data sheet of the appended data and populate the one-week returns using XLQ.

   2). Otherwise I paste from the SIP data to another worksheet the Price Change (4 weeks, 13 weeks, 26 weeks, and 52 weeks) and the % Rank Relative Strength (4 weeks, 13 weeks, 26 weeks, and 52 weeks)

   3). I sometimes will create a dependent variable from the average of these four periods.

      a). This weights the recent performance, as the 4 week is included in all four periods.

      b). The returns from longer time periods drop off entirely for more recent returns, i.e. there are no six-month returns in the data from four months ago.

   4). In Excel I create compound fields to facilitate the joins of data from a particular row to its dependent variable (returns). For example data from Alcoa for the 98th month of my data (3/29/09) needs to be joined to the 13 week Price Change from month 101 (6/30/09). The compound field would be 98AA (created in a column using the &).

   5). To do relative returns, I was taking the difference in return between the data row and the return of the Russell 3000 ETF (IWV). However, I found this gave distorted returns because the Russell 3000 is a weighted index and differed from the average of the SIP data.

   6). To analyze relative returns I use Excel pivots to find the average for each potential dependent variable, i.e. 13 week returns, that I can then use to subtract from the return for each row, i.e. 98AA for Alcoa the 98th month.

   7). Outliers

      a). Usually I will cap outlier returns, as the outliers have an excess impact on returns and are not likely to be replicated. I may change all losses below -50% to -50% and all gains above 100% to100%. Rarely will I do this to more than 2.5% of both the top and bottom of a dataset.

      b). The Rank Relative Strength (RRS) is a very nice way to handle the outlier problem. However, a stock with a RRS of say 85 thirteen weeks later may have had a RRS of 90 before the 13 weeks, meaning the returns were not so good. Would you use a change in RRS?)

   8). Worksheets for each time period and each type of dependent variable (absolute returns or relative returns) are created for import to Access to do the joins of returns for each data row.

d. In Access
   1). Import the necessary tables. I do some links, but the imported data processes faster.

   2). Create the relationships or joins to link the data with the appropriate consequent returns.

   3). Export the data back to Excel.

   4). (Excel look-up tables choke on this volume of data. The Access joins is much easier and faster.)

B. Analyze the data

1. I create a smaller file of selected fields to export to KnowledgeSEEKER, the data mining tool I use. I delete the compound fields used for the joins, the date column, the returns not to be used for the specific analysis, and other fields to be excluded from the specific anticipated analysis.
2. Import the data into KnowledgeSEEKER from the clipboard.
3. Partition the data. I partition the data 60-40 or 50-50. This is done at random by the tool. If the results from the learning set (say the 60% of records) don't compare well to the test set, the pattern is not uniform throughout the data and is not be trusted.
4. Adjust the settings
    a. Set precision for returns and standard deviation to .1.
    b. Level of significance depends upon the number of records. If I have 100,000 records after the partition I will set it to .0005. If I have a smaller number of records, say 20,000, I may set it to .02.
    c. Set other settings, such as Bonferonni and level of detail desired. I may need to set the edit to ignore certain fields.
5. Explore the trees looking for
    a. High count
    b. High returns
    c. Low standard deviation
    d. Returns not likely to be specific to a particular time period, i.e. based on price change 13 week, or all in two sectors.
    e. Consistency by month (or week)
    f. Consistency with the test partition
    g. Look at the actual records selected. Are they all in one industry? What do the charts look like?
6. Record the best looking results in a log outline format on another monitor.
    a. For each subset cell, calculate the count per month (or week) and the coefficient of variation. (SD/avg. This is the inverse of the Sharpe ratio, ignoring the "risk-free" rate of return.)
7. Go through the log making comparisons and select the best looking screens.
8. Do the screen variable make sense? Are they likely to persist?
9. Do a ranking comparison of the best looking screens.
    a. Go back to the Excel worksheet imported into KnowledgeSEEKER and exclude all columns and rows not in any of the best looking screens.
    b. Using groupings in pivot tables, create tables by month for each screen showing counts, average returns and standard deviations.
    c. Paste each pivot table into a worksheet and do a progressive series of rankings.

C. Apply the best screen or best two screens

1. Rate current screen stock selections using
    a. Create screens in SIP for the best one or two sets of variables.
        1). Some variables in SIP views are not available in SIP screens, such as Industry Price / Cashflow per Share. The filter has to be applied in Excel after the export.
        2). Backup the new user settings for use in restoring to future data sets.
    b. Import the stock list into TeleChart and rate each stock.
2. Start buying.

III. Findings along the way

A. It was hard to find consistent patterns.

1. The very high counts and relative returns from 10/08 through 08/09 were due to small caps doing much better than large caps and using a weighted index to define relative strength (Russell 3000, IWV).
2. When analyzing strong relative returns they were most often from just a few months.
3. There were fewer findings than when I previously did similar research. Does a market driven by panic obliterate the normal predictive patterns?

B. The effect of market timing is pervasive and overwhelms any other variable.

C. I haven't found an ideal way to handle outlier effects.

D. It is easy to confuse significance with consistency.

1. Significance relates to sampling reliability and validity.
2. Consistency relates to predictions.

E. I need to separate the timing variables (technical or price patterns) from the fundamental (longer-term). Variables are related to:

1. Fundamentals (sales, assets, earnings, debts, etc.)
2. Price patterns (i.e price as percent of 52-week high or technical indicators such as CCI, RSI & ATR). I want to use XLQ to collect indicator data and mine the data for timing clues.
3. Ratios between price and fundamentals (P/E, price to sales, price to book value, etc.) The ratios are the most productive.

F. One would expect results to be short-term when buying

1. Undervalued stocks.
2. Companies with poor fundamentals.
3. Based on reversals. While CANSLIM and most technical analysis affirms the value of momentum, the best results in this analysis were consistently in reversal patterns.