EXPLORATORY VALIDATION: WHAT WE CAN LEARN FROM TESTING INVESTMENT MODELS

Cem Kaner, J.D., Ph.D.
Florida Institute of Technology

STAR East
May 2011
ABSTRACT

Over the past couple of years, the airwaves have been flooded with commercials for investment-support software. Do your research with us, they promise, and you can make lots and lots of money in the stock market. How should we test such a product?

These products provide several capabilities: For example, they probably evaluate (estimate the value or direction of change of value) of the market as a whole or of individual stocks, and they provide trading strategies that tell you whether to buy, hold or sell a stock.

Most stock purchases and sales are fully automated—the trading programs buy and sell stocks based on their strategies. Every valuation rule and every strategy is a feature. We can test the implementation of these, but the greater risks lie in the accuracy of the models. If you execute the wrong trades perfectly, you reliably lose money. That’s not a happy-making feature.

Cem Kaner reports on work he’s been doing in this area for the last three years, presenting this as a case study of exploratory, high-volume test automation, done for the purpose of validation rather than verification.
**SOME POPULAR TRADING HEURISTICS**

- **Momentum**
  - "The trend is your friend (until it ends)"
  - "Buy high, sell higher"

- **Buy the dividend**
  - Buy a stock just before it pays a large dividend, collect the dividend and sell it.

- **Underbought, overbought**
  - "Buy low, sell high"

- **Buy out-of-favor stocks that are undervalued**
  - "Be greedy when others are fearful" (Buffett)
A heuristic

... turns into a strategy
when people start relying on it for investing

... turns into a feature
when it is incorporated into a program that
people use to decide whether to buy or sell stocks

Once it becomes a feature, testers should test it.

... but how?
My First Backtest

• I took control of my retirement savings in November 2008

• VectorVest was the first research service to attract my attention

• Today's comments are critical, but I enjoyed my interactions with VectorVest's staff, they have a good record with the Better Business Bureau (http://www.bbb.org/akron/business-reviews/investment-advisory-service/vector-vest-in-akron-oh-38000968) and I believe their product offers significant benefits, especially in the accessibility of its highly detailed historical fundamentals data.

Their product tour
http://www.vectorvest.com/freemovies/demo/vectorvestproducttour/vectorvestproducttour.html
VectorVest offers about 250 strategies, tailored for different expectations about market performance.

"VectorVest's Strategies and Searches"

http://www.youtube.com/watch?v=lb_h_mwKk-o
Consider a trading system with 250 strategies. Analyze the population every week to see which ones performed well. Should we rely on them?

If these were random performers, how many should perform better than the market?

At the 5% level of statistical significance – we expect 12.5 random strategies to look exceptional every week.

Each strategy is a candidate for more thorough testing.
"We believe that market direction is the single most important thing you need to know to consistently make money in the stock market. So consulting the Color Guard is the place to start before you do anything with your portfolio. That’s why we feature it prominently on our homepage and present it vividly for instant interpretation. Hopefully, it will serve you better now that we have The Color Guard: Clarified."

http://www.vectorvest.com/blog/page/The-Color-Guard-Clarified.aspx
The Color Guard

VectorVest U.S.

Stock Analysis and Portfolio Management System

Current as of Friday, July 18, 2008

The Color Guard is Neutral
VectorVest advocates caution when buying stocks at this time.
Click here for details...

VectorVest at a Glance

Updated as of 7/18/2008 for 8,369 stocks

<table>
<thead>
<tr>
<th>Buys</th>
<th>Sells</th>
<th>Holds</th>
</tr>
</thead>
<tbody>
<tr>
<td>810</td>
<td>3,939</td>
<td>3,620</td>
</tr>
<tr>
<td>(9.7%)</td>
<td>(47.1%)</td>
<td>(43.3%)</td>
</tr>
</tbody>
</table>

VVC Price: $25.94 BSR: 0.21
VVC RT: 0.88 MTI: 0.62

VectorVest Color Guard

<table>
<thead>
<tr>
<th>DATE</th>
<th>CGPRC</th>
<th>CGRT</th>
<th>CGBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/18/2008</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>7/17/2008</td>
<td>G</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>7/16/2008</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>7/15/2008</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>7/14/2008</td>
<td>R</td>
<td>R</td>
<td>Y</td>
</tr>
<tr>
<td>7/11/2008</td>
<td>R</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Market Timing Indicator

From 1/18/2008 to 7/18/2008

Copyright © Cem Kaner
• The color system summarizes trends over the past few days / weeks.
• According to VectorVest:
  – yellows mean there is no trend to follow,
  – red means the market is declining and you should NOT buy anything tomorrow, and
  – green means the market is rising and you feel good about buying.
To study the ColorGuard system as a predictor of the market, I downloaded Standard and Poors' S&P-500 index prices from January 4, 1999 through early Sept 2009.

I then computed percentage price changes:

• percent gain or loss in the S&P compared to the current day
• percent gain or loss between the current day value and the value 5 trading days from the current day.
• after 15 trading days
• after 30 trading days.

I also looked at 2-day, 3-day and 4-day for some analyses, but the results were the same as 1-day and 5-day so I stopped bothering.

• The average day-to-day change in the market was 0.0027% (flat over 10 years)
RESULTS (THE DATA)

From 1999 to September 2009:

- 289 trading days rated **GREEN / GREEN / GREEN (GGG)** (buy)
- 285 rated **RED / RED / RED (don’t buy)**.
- After **GGG** days, S&P index went **DOWN** an average of
  - **-0.05%** the day after a **GGG** rating,
  - **-0.24%** five days after,
  - **-0.14%** 15 days after, and
  - **-0.33%** 30 days after.
- After **RRR** days, S&P went **UP** an average of
  - **0.24%** the day after an **RRR**,
  - **0.29%** five days after,
  - **0.42%** 15 days after and
  - **1.08%** in the 30 days after an **RRR**

Copyright © Cem Kaner
RESULTS

Average profit per trade (in %)

Sell N days after purchase

- Buy on GREEN day
- Buy on RED day
ANOTHER APPROACH THAT DOESN'T SEEM VERY PROMISING

Buy a stock when it hits its 52-week high
MOMENTUM: 52-WEEK HIGH?

There is some evidence that stocks bought at a 52-week high are likely to be higher 6 months later.

- The 52-Week High and Momentum Investing
  Thomas. J. George & Chuan-Yang Hwang
  http://www.bauer.uh.edu/TGeorge/papers/gh4-paper.pdf

- Is the 52-week high momentum strategy profitable outside the US?
  Ben R. Marshall & Rachael M. Cahan
  http://www.ingentaconnect.com/content/routledg/rafe/2005/00000015/00000018/art00002

- 52 Week High and Momentum Investing: A Partial Replication of George and Hwang's Results
  Bill Jannen & Vincent Pham
52-WEEK HIGH

52-week high is also recommended (often) as a short-term trading strategy:

"The 52-Week High Breakout Strategy attempts to buy a stock when the current day’s closing price reaches a 52-week high.

The stock’s 10-day average trade volume must also be higher than the stock’s 90-day average trade volume to generate a trade signal.

The strategy exits the position after hitting a Profit Target (with a 10% default), or after a specific period (with a default of ten days)."

http://erresearch.fidelity.com/backtesting/viewstrategy?category=Breakout&wealthScriptType=FiftyTwoWeekHigh
52-WEEK HIGH

I found similar ideas on VectorVest, Zachs, TD Ameritrade, and several other investment sites.

How can we test this?

• A strategy tells us
  – When to buy,
  – How much to buy,
  – When to sell.
• In backtesting, we apply the strategy to historical data
BACKTESTING 52-WEEK HIGH

• The provided strategies often included constraints that didn't appear to have been dictated by the underlying idea, such as:
  • Prices restricted to a narrow range (between $x$ and $y$)
  • Carefully picked trailing stops.
• In 2008/2009, I replicated several claimed backtests, using their parameters over their published time periods
  • I usually obtained the claimed results
  • But when I changed parameters, the results were often much worse.

WHY?
What's our starting point?

Using Fidelity's Wealth-Lab Pro

- Buy the morning after the 52-week high
- Buy stocks ≥ $1.00
- Sell at the end of the day
- (If the stock is still rising, we should see an increasing trend.)

The following data:

- 5 years (4/2006 to 3/2011)
- S&P 100 stocks (NASDAQ 100, same results)

For a free trial of Wealth-Lab Pro, go to

1-DAY RESULTS
2-DAY RESULTS

![Graph showing profit distribution of strategy trades. The graph displays the number of trades at different profit percentages. The majority of trades are near zero profit, with a significant number of trades in the positive profit range. The most common profit percentages are 0%, 5%, and 10%. The graph also includes a bar chart with the number of trades at each profit percentage, indicating a skewed distribution towards positive profits.](image-url)
The basic results are not promising

But when 1158 of 2782 trades are profitable, we might look for ways to separate the 1158 winners from the 1624 losers.

– For example, if we restrict our attention to companies that are debt-free, over the 5 years, we would make 36 winning trades and only 29 losers.

OPTIMIZATION is the process of adding or changing parameters to our backtests to:

- Gain more per trade or
- Increase the percentage of profitable trades
OPTIMIZATION & OVERFITTING

• OPTIMIZATION
  – Adding or changing parameters to our backtests to:
    ▪ Gain more per trade or
    ▪ Increase the percentage of profitable trades

• OVERFITTING
  – It's easy to over-optimize, obtaining results that are limited to the specific data set that was tested.
  – Good results from backtesting are encouraging, but need further testing to control the risk of overfitting (and other risks).
Buy the dividend

- Many companies pay dividends (money to the shareholder, every month, quarter, or year)
- Buy a stock for $20
  - get a 50¢ dividend each quarter
  - This is a $2 (10%) annual interest rate
- The day after the dividend, the stock price drops by the dividend amount

But when we buy a company, we are buying a future income stream

- So, if everyone has high confidence in the company, prices rise back, in anticipation of future profits.
$25 exchange traded note, 37.5¢ (6%) dividend.

Copyright © Cem Kaner
BCSPRA $25 stock, 44.37¢ (7.1%) dividend.
BUY THE DIVIDEND

• This is such an obvious way to make money…
• It is such a cool strategy that several ETF's specialize in it.
• Obviously, everyone should use it.

• So … we study it in the investment modeling class' first assignment
Uh-Oh ... Capital Product Partners

CPLP $7-11 stock, variable dividend, currently about 11%.
BACKTESTING RESULTS

• The basic buy-the-dividend (just before the dividend date) strategy fails badly

• Prices drop by the amount of the dividend when the dividend is paid and may take weeks (or longer) to recover

• Some identifiable stocks (e.g. CPLP and other high-dividend energy stocks) are better for sell-the-dividend

• Buy-the-dividend appears profitable (but not amazingly profitable) if you buy a month ahead of the large dividend instead of a few days ahead

• Buy-the-dividend might be profitable a few days ahead on modest dividends (e.g. 4% rather than 8%), but brokers' commissions can play a big role
One more example

Intense backtesting on:
- NASDAQ,
- Dow Jones and
- S&P 500

up to 2004
**Representative Finding**

![Bar Graph]

**S&P 500**

3 Days in a Row

- Average One-Week Return of the S&P 500 after the S&P 500 rises for at least three days in a row
- Average One-Week Return of the S&P 500 after the S&P 500 falls for at least three days in a row
Replications & Extensions

Short Term Trading Strategies That Work
Larry Connors
Cesar Alvarez

High Probability ETF Trading
7 Professional Strategies to Improve Your ETF Trading
Larry Connors
Cesar Alvarez
Connors Research

Copyright © Cem Kaner 32
JEFF FULLER (AT FLORIDA TECH):

- Replicated Connors' & Alvarez's results (*High Probability ETF Trading*)
- Extended the study to more ETF's
- Extended the method to add an *extra aggressive* variant
- "For my implementation of an 'extra aggressive' version of each strategy, when a 'second unit' is mentioned in the aggressive version, I simply considered 'another unit.' By doing this it means that every time the condition of the aggressive version is met another unit is bought, rather that capping out at two units. This results in more capital invested but, as will be shown, also results in significantly higher win percentage as well as a higher average percent profit."
EXAMPLE STRATEGY

The long method is described in *High Probability ETF Trading*:

1. Today the ETF is above the 200-day moving average.
2. Today the ETF closes below its 5-day moving average.
3. Two days ago the high and low price of the day is below the previous day’s high and low.
4. Yesterday the high and low price of the day is below the previous day’s.
5. Today’s high and low price is below yesterday’s.
6. Buy on the close today.
7. Aggressive Version – Buy a second unit if prices close lower than your initial entry price anytime you’re in the position.
8. Exit on the close when the ETF closes above its 5-day simple moving average.
# Example Results ( Fuller )

<table>
<thead>
<tr>
<th>Stock</th>
<th>Trades</th>
<th>Avg % P/L</th>
<th>Avg Days Held</th>
<th>Win %</th>
<th>Avg Buys In</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIA</td>
<td>50</td>
<td>0.75</td>
<td>3.3</td>
<td>86.0</td>
<td>2.7</td>
</tr>
<tr>
<td>EEM</td>
<td>31</td>
<td>1.66</td>
<td>3.2</td>
<td>93.5</td>
<td>2.5</td>
</tr>
<tr>
<td>EFA</td>
<td>25</td>
<td>1.17</td>
<td>3.2</td>
<td>92.0</td>
<td>2.4</td>
</tr>
<tr>
<td>EWH</td>
<td>48</td>
<td>1.65</td>
<td>3.8</td>
<td>87.5</td>
<td>2.8</td>
</tr>
<tr>
<td>EWJ</td>
<td>32</td>
<td>1.32</td>
<td>3.6</td>
<td>84.4</td>
<td>2.9</td>
</tr>
<tr>
<td>EWT</td>
<td>27</td>
<td>1.30</td>
<td>3.9</td>
<td>88.9</td>
<td>3.0</td>
</tr>
<tr>
<td>EWZ</td>
<td>44</td>
<td>2.52</td>
<td>2.6</td>
<td>97.7</td>
<td>2.0</td>
</tr>
<tr>
<td>FXI</td>
<td>17</td>
<td>2.06</td>
<td>3.4</td>
<td>82.4</td>
<td>2.6</td>
</tr>
<tr>
<td>GLD</td>
<td>18</td>
<td>1.39</td>
<td>4.0</td>
<td>94.4</td>
<td>2.8</td>
</tr>
<tr>
<td>ILF</td>
<td>36</td>
<td>2.40</td>
<td>2.6</td>
<td>97.2</td>
<td>1.8</td>
</tr>
<tr>
<td>IWM</td>
<td>40</td>
<td>1.12</td>
<td>3.4</td>
<td>100.0</td>
<td>3.0</td>
</tr>
<tr>
<td>IYR</td>
<td>40</td>
<td>0.80</td>
<td>3.6</td>
<td>90.0</td>
<td>2.9</td>
</tr>
<tr>
<td>QQQQ</td>
<td>42</td>
<td>1.81</td>
<td>3.3</td>
<td>90.5</td>
<td>2.9</td>
</tr>
<tr>
<td>SPY</td>
<td>66</td>
<td>1.22</td>
<td>2.6</td>
<td>95.5</td>
<td>2.1</td>
</tr>
<tr>
<td>XHB</td>
<td>4</td>
<td>1.61</td>
<td>3.5</td>
<td>75.0</td>
<td>3.3</td>
</tr>
<tr>
<td>XLB</td>
<td>39</td>
<td>0.87</td>
<td>3.9</td>
<td>84.6</td>
<td>3.4</td>
</tr>
<tr>
<td>XLE</td>
<td>51</td>
<td>1.45</td>
<td>3.5</td>
<td>90.2</td>
<td>3.0</td>
</tr>
<tr>
<td>XLF</td>
<td>34</td>
<td>0.86</td>
<td>3.4</td>
<td>94.1</td>
<td>2.8</td>
</tr>
<tr>
<td>XLI</td>
<td>36</td>
<td>1.03</td>
<td>3.5</td>
<td>91.7</td>
<td>2.8</td>
</tr>
<tr>
<td>XLV</td>
<td>44</td>
<td>0.77</td>
<td>3.0</td>
<td>88.6</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Model Testing

- Each buy/sell rule reflects an underlying model of the market’s historical behavior.
- But in an “investment program”, each implemented strategy is implemented as a feature of the program.
- So if we are testing an investment program (financial services are now a huge part of our economy), we are testing each strategy.
## Assessing Models: 7 Risks

<table>
<thead>
<tr>
<th>Model risk</th>
<th>The model is theoretically or empirically incorrect.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characterization risk</td>
<td>The model is correct, but the description of it (the spec) has errors.</td>
</tr>
<tr>
<td>Comprehension risk</td>
<td>We misunderstand the model. Our code accurately implements the wrong model</td>
</tr>
<tr>
<td>Implementation risk</td>
<td>Coding errors. Data storage / retrieval errors. Our code inaccurately reflects our intent.</td>
</tr>
<tr>
<td>Execution / environmental risk</td>
<td>We attempt the correct trade but the software / hardware platform is too slow, can't handle the data volume, the data feed is too slow, etc. As a result, we fail in our efforts to buy or sell at the desired price.</td>
</tr>
<tr>
<td>Tool risk</td>
<td>Our test tool misleads us by corrupting the software under test or by missing failures or by giving us false alarms</td>
</tr>
<tr>
<td>Scope risk</td>
<td>Our model is properly developed but is not appropriate to today's circumstances. Can we recognize when market activity is out of scope of our model?</td>
</tr>
<tr>
<td>Verification</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---</td>
</tr>
<tr>
<td><strong>Model risk</strong></td>
<td>The model is theoretically or empirically incorrect.</td>
</tr>
<tr>
<td><strong>Characterization risk</strong></td>
<td>The model is correct, but the description of it (the spec) has errors.</td>
</tr>
<tr>
<td><strong>Comprehension risk</strong></td>
<td>We misunderstand the model. Our code accurately implements the wrong model.</td>
</tr>
<tr>
<td><strong>Implementation risk</strong></td>
<td>Coding errors. Data storage / retrieval errors. Our code inaccurately reflects our intent.</td>
</tr>
<tr>
<td><strong>Execution / environmental risk</strong></td>
<td>We attempt the correct trade but the software / hardware platform is too slow, can't handle the data volume, the data feed is too slow, etc. As a result, we fail in our efforts to buy or sell at the desired price.</td>
</tr>
<tr>
<td><strong>Tool risk</strong></td>
<td>Our test tool misleads us by corrupting the software under test or by missing failures or by giving us false alarms.</td>
</tr>
<tr>
<td><strong>Scope risk</strong></td>
<td>Our model is properly developed but is not appropriate to today's circumstances. Can we recognize when market activity is out of scope of our model?</td>
</tr>
<tr>
<td>Risk Type</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Model risk</td>
<td>The model is theoretically or empirically incorrect.</td>
</tr>
<tr>
<td>Characterization risk</td>
<td>The model is correct, but the description of it (the spec) has errors.</td>
</tr>
<tr>
<td>Comprehension risk</td>
<td>We misunderstand the model. Our code accurately implements the wrong model.</td>
</tr>
<tr>
<td>Implementation risk</td>
<td>Coding errors. Data storage / retrieval errors. Our code inaccurately reflects our intent.</td>
</tr>
<tr>
<td>Execution / environmental risk</td>
<td>We attempt the correct trade but the software / hardware platform is too slow, can't handle the data volume, the data feed is too slow, etc. As a result, we fail in our efforts to buy or sell at the desired price.</td>
</tr>
<tr>
<td>Tool risk</td>
<td>Our test tool misleads us by corrupting the software under test or by missing failures or by giving us false alarms.</td>
</tr>
<tr>
<td>Scope risk</td>
<td>Our model is properly developed but is not appropriate to today’s circumstances. Can we recognize when market activity is out of scope of our model?</td>
</tr>
</tbody>
</table>
Backtesting

• Select a set of stocks, to study over a period of time
• Select attributes of:
  – The economy
  – The behavior of the market as a whole
  – The underlying strength of the company
  – The price of the stock
• Decide what combinations of these attributes should trigger a buy or sell
• For each stock in the set, for each day in the period
  – Buy or sell according to the attribute values
We specify whether to buy or sell (and how)
Many types of conditions, can control many separate buy/sell rules
We can use pre-built attributes or custom ones.

Wealth-Lab builds complex rules by given you a very complex wizard. But if you want to add your own attributes:

- View the underlying strategy implementation
- Write code to add new attributes

All of this testing is:

- Intensely automated, but
- Non-repetitive
Software testing?

• Most trades are automated
  – Buy/sell decisions come from models, implemented in trading programs
  – Testing this software is extremely important
  – Trillions of dollars of trades are automated
The testing that I am illustrating is exploratory:

- No regression tests
- Intensely automated
- Quasi-experimental designs
- Carefully planned
- Tightly controlled
- Theory-guided
- Complex issues
- Difficult oracles
- No regression tests
Skilled black-box testers have done exploratory validation at every company I have worked for or consulted to.

Exploratory testing blends:

- Knowledge and skill in many generic test techniques
- Knowledge of the application under test and its quality-related risks. **Many of these are not related to coding errors.**
- Knowledge of the test techniques / technology of the application's field.
Cem Kaner has pursued a multidisciplinary career centered on the theme of the satisfaction and safety of software customers and software-related workers.

With a law degree, a doctorate in Experimental Psychology, and seventeen years in Silicon Valley, Cem joined Florida Institute of Technology in 2000 as Professor of Software Engineering. His research focuses on the question, How can we foster the next generation of leaders in software testing?

Cem is the senior author of Testing Computer Software (with Jack Falk and Hung Quoc Nguyen), Bad Software (with David Pels), and Lessons Learned in Software Testing (with James Bach and Bret Pettichord).

The Association for Computing Machinery recently honored Cem with its Making a Difference Award