Understanding Mutual Fund Success Using Boosted Tree Models

Noah Berg and Research Mentor Dr. Jeong Ho (John) Kim
Florida State University, College of Business

Objective

The objective of our data analysis is to determine which characteristics of mutual funds are most impactful for mutual fund success. While novel research in the field suggests that mutual fund success is based on fund characteristics, our research seeks to challenge these assertions and provide evidence it is truly stock characteristics that impact fund success the most.

Introduction and Background Information

It's important to understand what a mutual fund is and how they operate. Mutual funds are the collective invested assets of investors into a single pool. This means when the value of the mutual fund goes down, all investors lose that percentage of value as well. Mutual funds have become a popular choice for both large scale Wall Street investments and individuals looking to diversify their personal investments. With the age of AI on the horizon it is vital more than ever we use advance data analysis techniques to process the large amount of data collected on mutual funds. Our data evaluates 3232 starting from 1997 to 2017 analyzing over 50 variables.

A vital aspect to understanding mutual funds is what 'success' even means. Just like any other business, the fund managers want to maximize their investment to generate as much profit as possible. For our research, we use excess rate of return and capital asset pricing model index to evaluate how well a mutual fund will outperform the general market.

Finally, we must understand mutual fund characteristics and the difference between stock and fund variables. . First are fund characteristics which are variables that describe the fund itself. Some examples of fund characteristics fund size (number of investors) or the periodic change in total assets in a fund. The second type of factors that researcher's study are stock characteristics. These are variables that describe the actual value of the securities of the stock.

Methods

We used Boosted Regression Trees (BRTs) to analyze correlations between characteristics (our X variables) and fund performance (our Y variables). We chose to use BRTs have exhibited strong predictive performance in various fields and can handle large, high-dimensional data sets, because they perform both variable selection and shrinkage in an automated fashion.

The following are the steps used for the model:

1) Establish the regression trees.

$$f(x) = \sum_{j=1}^{J} c_j I\{x \in S_j\},$$

2) To boost take the sum of the regression trees.

$$f_B(x) = \sum_{b=1}^{B} \mathcal{T}_b\left(x; \{S_{b,j}, c_{b,j}\}_{j=1}^{J}\right),$$

3) Split based on variables

$$\begin{aligned}\hat{c}_{j,b} &= \min_{c_{j,b}} \sum_{x_t \in S_{j,b}} \left[y_{t+1} - (f_{b-1}(x_t) + c_{j,b})\right]^2 \\ &= \min_{c_{j,b}} \sum_{x_t \in S_{j,b}} \left[e_{t+1,b-1} - c_{j,b}\right]^2,\end{aligned}$$

4) Sum the reductions in empirical errors

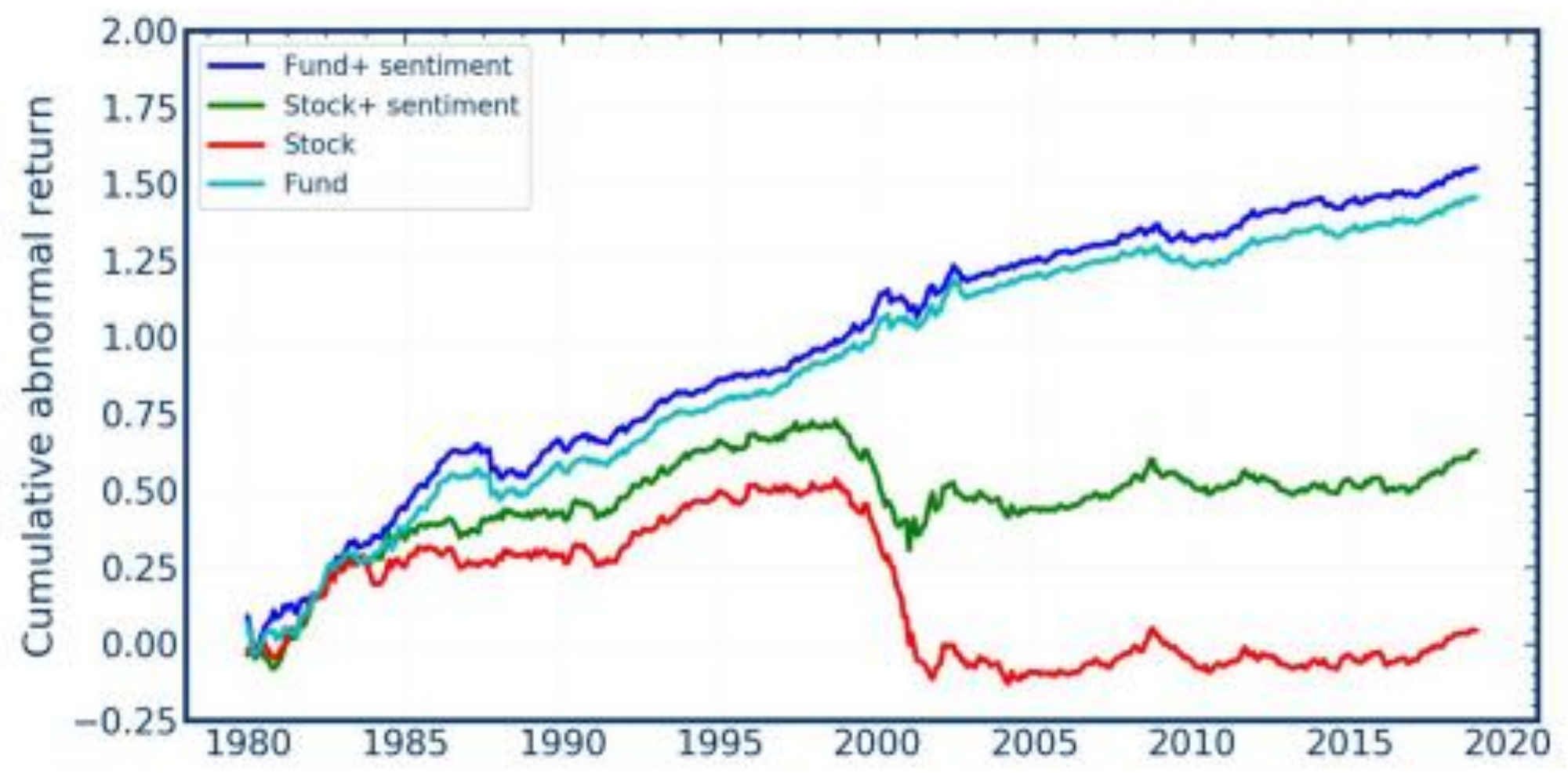
$$I_l(\mathcal{T}) = \sum_{j=2}^{J} \Delta e(j)^2 I(x(j) = l),$$

$$f_b(x) = f_{b-1}(x) + \lambda \sum_{j=1}^{J} c_{j,b} I\{x \in S_{j,b}\}.$$

5) Use a shrinkage parameter to determine how much each boosting iteration contributes to overall fit.

This method has more intract details references on pages 14-21 of the Li and Rossi paper. Using this method, we coded it into a statistical analysis program known as R which analyzes the data. This statistical software is the most capable at processing the large amount of data needed for this research project.

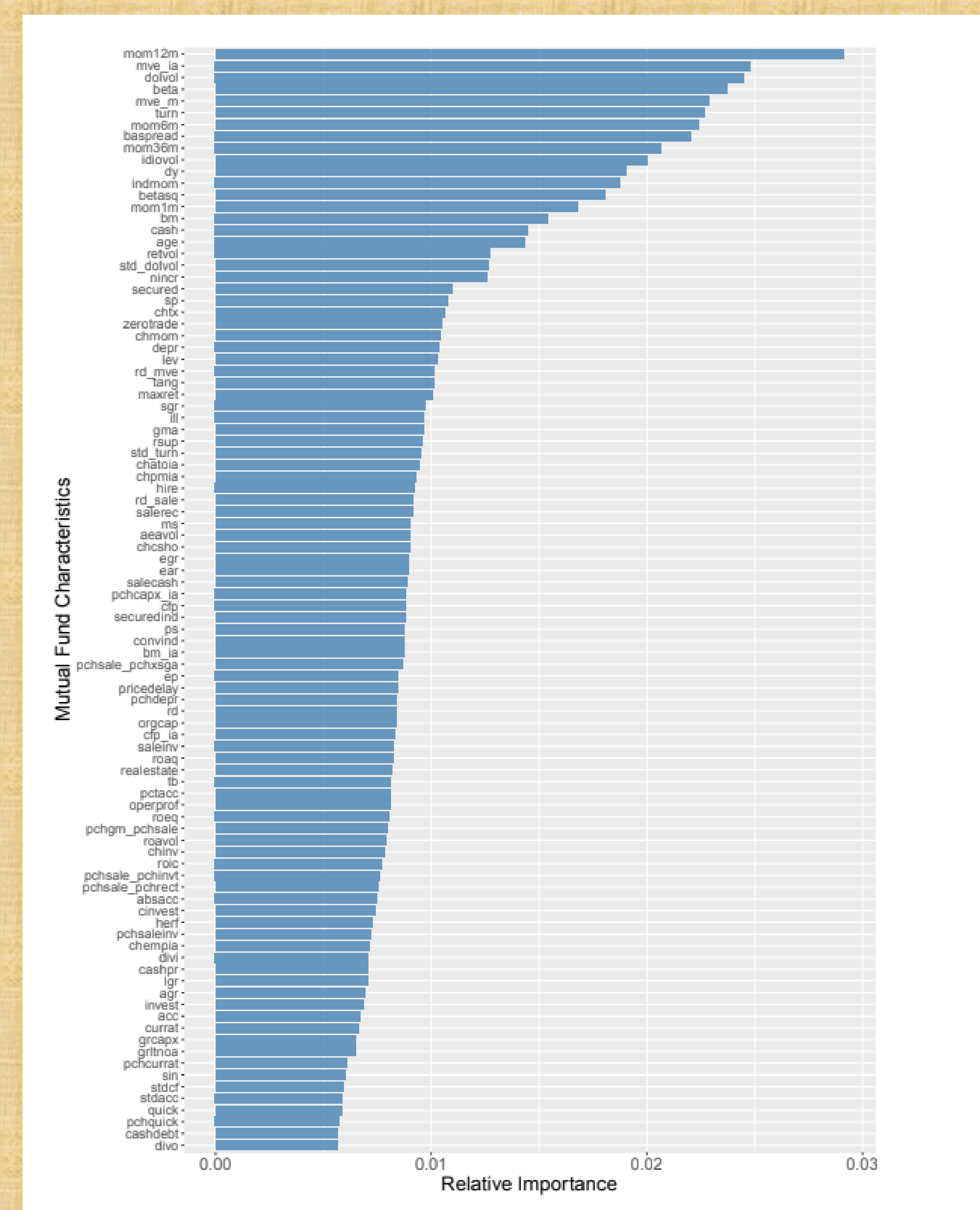
Figure A.3: Cumulative abnormal returns of equally-weighted long-short prediction portfolios.

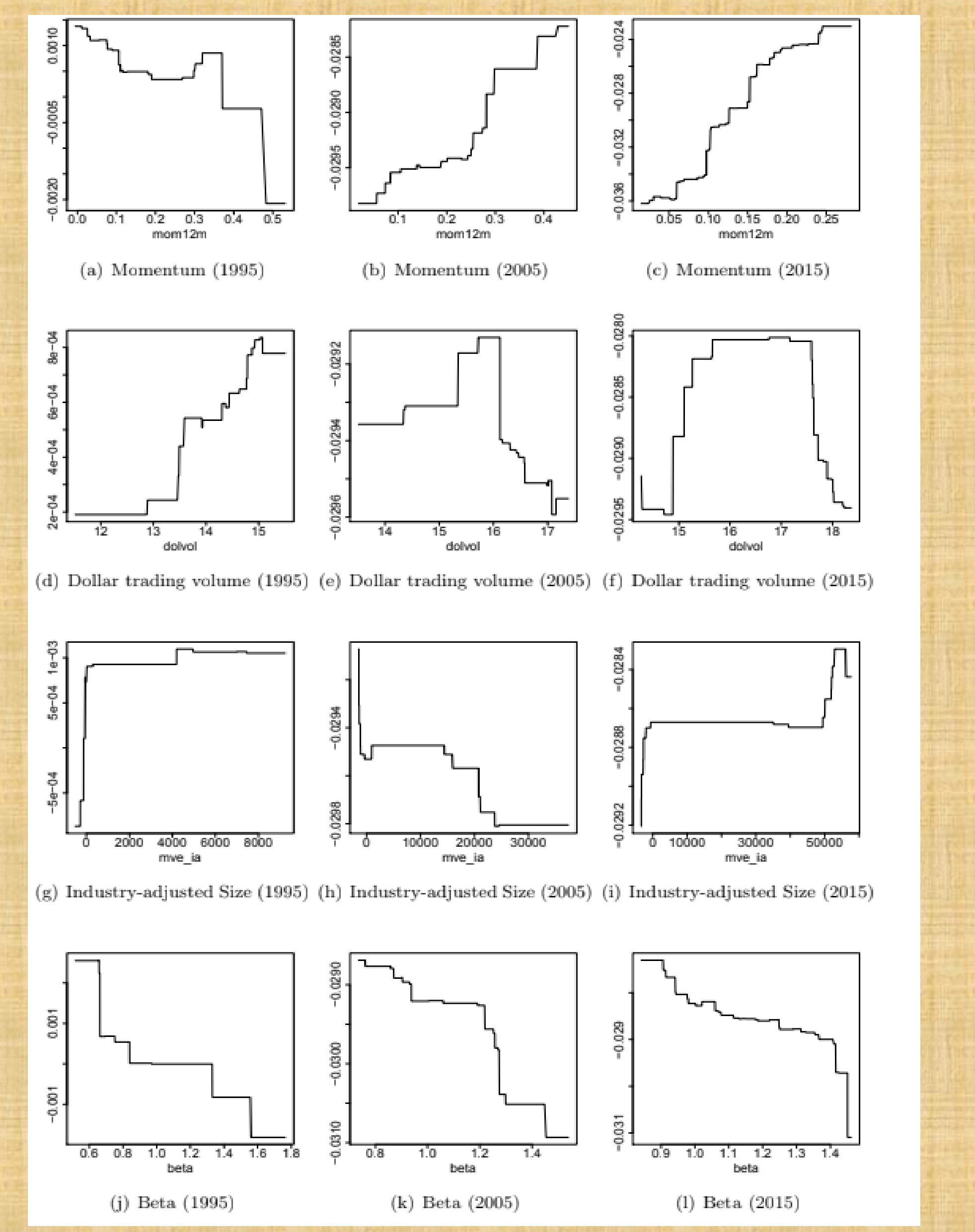


This figure plots the cumulative abnormal returns of equally-weighted long-short decile portfolios that use different information sets to predict abnormal returns. We consider fund-specific and stock-specific characteristics combined with sentiment.

Conclusion

So far the results of our analysis have came up as inconclusive. With more time to clean the data collected from Morningstar, not associated with this board, and additional time to data process it would returned an value determining correlations between variables. Below is data collected in the Li and Rossi paper which used a different method, but represents the type of data we hope to collect. Below the method sections is the fund versus stock data chart we hope to recreate.





(a) Momentum (1995) (b) Momentum (2005) (c) Momentum (2015)

(d) Dollar trading volume (1995) (e) Dollar trading volume (2005) (f) Dollar trading volume (2015)

(g) Industry-adjusted Size (1995) (h) Industry-adjusted Size (2005) (i) Industry-adjusted Size (2015)

(j) Beta (1995) (k) Beta (2005) (l) Beta (2015)

References

Barber, Brad M., et al. "Which factors matter to investors? evidence from mutual fund flows." *Review of Financial Studies*, vol. 29, no. 10, 21 June 2016, pp. 2600–2642, https://doi.org/10.1093/rfs/hhw054.

"Boosted Tree Regression in R." *KoalaTea*, KoalaTea, 14 Dec. 2023, koalatea.io/r-boosted-tree-regression/.

Freyberger, Joachim, and Michael Weber. "Dissecting characteristics nonparametrically." *SSRN Electronic Journal*, 2017, https://doi.org/10.2139/ssrn.2820700.

Hong, Harrison, and Jeremy C. Stein. "Chapter 14. A unified theory of underreaction, momentum trading, and overreaction in asset markets." *Advances in Behavioral Finance, Volume II*, 31 Dec. 2005, pp. 502–540, https://doi.org/10.1515/9781400829125-017.

Kaniel, Ron, et al. "Machine-learning the skill of mutual fund managers." *SSRN Electronic Journal*, 6 Apr. 2023, https://doi.org/10.2139/ssrn.4028339.

Li, Bin, and Alberto G. Rossi. "Selecting mutual funds from the stocks they hold: A machine learning approach." *SSRN Electronic Journal*, 6 Dec. 2020, https://doi.org/10.2139/ssrn.3737667.

Zach. "Lasso Regression in R (Step-by-Step)." *Statology*, 13 Nov. 2020, www.statology.org/lasso-regression-in-r/.