Revisiting the Nominal Share Price Puzzle: A Machine Learning Approach*

Jialie Chen† Cong Feng‡ Hassan Ilyas§ Gaurav Kankanhalli¶

April 3, 2019

Abstract

We reexamine the nominal share price puzzle (Weld et al., 2009) using a novel, proprietary dataset containing detailed information on 500,000 individual investors' trading activities across all mutual funds in a large family of funds from 2000-2015. Notably, three funds belonging to the fund family conduct stock splits during the sample period, allowing us to investigate the effects of splits on investors' trading behavior. In order to estimate the heterogeneous effects of splits on investors with varying degrees of sophistication, we propose a machine learning algorithm to classify investors into sophisticated and unsophisticated groups. Our algorithm, built on Gradient Boosted Machines, leverages the large feature space consisting of investors' attributes derived from their past trading history and demographic information to provide robust out-of-sample predictions of investor sophistication. Using this classification, we address the key question of whether splits attract abnormal inflows from unsophisticated investors, and whether these effects are short-lived. We then trace the responses of sophisticated investors to splits. Our investor-level data and machine learning approach allow us to overcome fundamental limitations of prior studies on the effects of splits arising from their use of market-level data and empirical methodologies designed to measure homogeneous investor effects.

*We thank Maureen O'Hara and Gideon Saar for valuable comments. All errors remain our own.
†Sam M. Walton College of Business, University of Arkansas. Email: JChen@walton.uark.edu.
‡School of Business Administration, University of Mississippi. Email: cfeng@bus.olemiss.edu.
§SC Johnson Graduate School of Management, Cornell University. Email: hi75@cornell.edu.
¶SC Johnson Graduate School of Management, Cornell University. Email: gak79@cornell.edu.

It has been observed that firms in the United States (and several other major economies) have, since the 1930s, undertaken policies with the aim of maintaining a constant nominal share price. This phenomenon has been dubbed the nominal share price puzzle in reference to the fact that over the same period, the nominal prices for all other types of goods in the economy have risen, on average, tenfold (Weld, Michaely, Thaler, and Benartzi, 2009). That firms actively split their stocks to maintain their stock prices at a constant nominal value presents a conundrum as splits are associated with costs for both firms and investors. Notably, splits increase investors' trading costs, which are typically charged on a per share basis. Further, splits increase the effective transaction costs per share due to an increase in the proportional bid-ask spread, which is bounded from below by the prevailing tick size. In addition to the direct costs of splitting incurred by firms, there is some evidence that splits are followed by increased return volatility (Ohlson and Penman, 1985). In this proposal, we aim to study investor responses to splits and shed light on the underlying economic mechanisms driving this prevalent, yet little-understood, firm policy.[1] We will do so using data and methods which directly address fundamental limitations of prior work on this topic. First, we introduce a novel investor-level dataset that allows us to disentangle the confounded and endogenous theoretical explanations. Second, we develop a machine learning-based investor classification algorithm, which enables us to study the heterogenous effects of splits on various classes of investors.

We conjecture that the failure of existing research to clarify the underlying economic mechanism for splits arises from two fundamental challenges. The first is an issue of identification — given that splits are endogenous, empirically disentangling the factors driving firms' decisions to split, and its effect on investor behavior, is difficult. The second challenge arises from the fact that splits may have heterogenous effects on different groups of investors. For instance, it can be argued that unsophisticated investors are likely to respond most positively to splits, while sophisticated investors may either be neutral to splits or may trade in a manner that offsets the changes induced by the inflow of unsophisticated investors. Therefore, analyzing market-wide variables will obscure the true effects of splits. Prior studies have only been able to focus on aggregate outcomes due to their use of market-level datasets like CRSP or TAQ. We propose a novel setting and machine-learning methodology that resolve both these issues.

Our study uses the setting of mutual fund splits. Mutual fund splits are an ideal context in which to revisit the nominal share price puzzle. First, arguably only one explanation for the puzzle, the marketability explanation, applies to mutual funds. This greatly simplifies the identification challenge. Since mutual funds fees and trading costs are based upon the investment amount, and not units traded, the pay-to-play explanation does not apply. The signaling explanation is unlikely to hold for mutual fund splits as these splits are less costly than common stock splits and, unlike firm managers, fund managers do not have the need to convey information on financial prospects to investors.[2]

Our study utilizes a novel, proprietary dataset comprising the complete trading activity and rich demographic information of a random sample of 500,000 investors across all funds belonging to one of the largest mutual fund families in China (with over $130 billion in assets under management) from 2000-2015. Crucially, three funds in the family conducted splits during this period such that the price of a share post-split was reset to 1 CNY. The granular information on investor characteristics in our data will allow us to tease out heterogeneous effects of these splits on investors of varying degrees of sophistication.

In order to classify investors as sophisticated or unsophisticated we plan to consider a detailed set of attributes. These include age, gender, location, time since account opening (proxy for investing experience), income (proxied by total investment size), number of trades, number of funds invested in, turnover, volatility of returns, and drawdowns. Classifying investor sophistication using traditional methods is challenging for at least two reasons. First, the feature space is large and includes variables of all types such as binary, categorical, and numerical. Second, the ways in which these features influence investor sophistication can be highly non-linear with several levels of interactions, and potentially complex hierarchical relationships (Easley, Lopez de Prado, O'Hara, and Zhang, 2019). For instance, prior studies have individually found that the investors' trading performance is affected by their gender, trading frequency, attention to fees, or whether they trade over the phone or online (Barber and Odean, 2000, 2001, 2002; Barber, Odean, and Zheng, 2005). However,

[1] Weld et al. (2009) propose four broad explanations. First, the "marketability explanation" is that splits enable firms to keep their stock price in a certain range in order to increase their stock's appeal to smaller, less wealthy investors. This would lead to a more diverse ownership base, and may improve their stock's liquidity. The empirical evidence on this explanation is mixed with some studies finding positive liquidity effects (Schultz, 2000), and others finding such effects, if they exist, are small and short-lived (Lakonishok and Lev, 1987; Byun and Rozeff, 2003). Second, the "pay-to-play explanation" is that firms split in order to encourage brokers and dealers to promote their stock and engage in market making by allowing them to maintain high relative bid-ask spreads (and thus profits). However, this explanation is inconsistent with the fact that the markets have seen multiple drops in the tick size over the last century but nominal share prices have remained constant. A third "signaling explanation" argues that firms engage in stock splits as a costly signal of their financial strength to investors. However, there is no clear evidence that splitting firms indeed perform better post-split. Moreover, Easley, O'Hara, and Saar (2001) find no reduction in information asymmetry after splits. These findings jointly cast doubt on the efficacy of stock splits as signals. In the absence of compelling evidence in favor of any of the three explanations, Weld et al. (2009) conclude in favor of the residual explanation, that firms' decisions to maintain a constant share price is a result of "norms."

[2] A few prior studies have investigated mutual fund splits (see, for example, Rozeff, 1998 and Fernando, Krishnamurthy, and Spindt, 1999). They provide suggestive evidence in favor of the marketability hypothesis. However, both studies use aggregate mutual fund flows data and consequently are unable to overcome the two fundamental challenges we have outlined.

attempting to classify the sophistication level of a male investor who trades infrequently and only via the internet, for instance, is a non-trivial problem under usual regression-based approaches due to the need to consider all interactions of these attributes. Crucially, the lack of a robust prior about the potentially complex ways in which these interactions work together (and change over time) to influence investor sophistication renders traditional approaches, in which the researcher must specify the relationship *ex ante*, unsuitable. Therefore, we resort to machine learning algorithms that offer greater flexibility and impose fewer restrictions on the underlying data generating process.

Specifically, we aim to tackle this classification problem by using Gradient Boosted Machines (GBMs). GBMs sequentially build an ensemble of weak and shallow trees where each tree learns and improves on the previous one (Friedman, 2001; Chen and Guestrin, 2016; Ke et al., 2017; Biau et al., 2018). The combination of several of these weak trees produces a powerful "committee" that has been show to have superior predictive abilities (for examples see Hutchinson et al., 2011; Pittman and Brown, 2011; Johnson and Zhang, 2014). Moreover, GBMs provide substantial flexibility as the researcher has several loss functions and hyper-parameters that can be tuned to improve predictive ability. Crucially for the investor classification problem, GBMs tend to perform well even when the feature space contains a mix of categorical and numerical variables, and also when data is missing.

Our approach will address two common pitfalls of machine learning algorithms: (1) Overfitting, and (2) Lack of interpretability. We will avoid overfitting by employing more sophisticated versions of GBMs that use only a subsample of the data to implement stochastic gradient descent or add regularization terms to the objective function (Friedman, 2002; Rashmi and Gilad-Bachrach, 2015; Chen and Guestrin, 2016). Regarding lack of interpretability, an important advantage of using GBMs is that their interpretability can be enhanced by using tools such as variables importance (see Easley et al., 2019 for an example in the context of random forests), Accumulated Local Effects plots (Apley, 2016), and Local Interpretable Model-Agnostic Explanations (Ribeiro, Singh, and Guestrin, 2016). Using these tools we can infer which investor characteristics seem to predict their performance. In doing so, we hope to contribute not only to the understanding of mutual fund splits, but also to the rich literature on the determinants of investor performance by jointly considering a large number of attributes.[3]

Our investor classification algorithm will involve the following key steps. First, we will utilize the trading data to calculate the positions and returns for all individuals across all days and funds. We will calculate risk-adjusted performance as each investor's alpha relative to reasonable benchmark factors. Second, we will generate the feature space consisting of demographic information (gender, age, investing experience) as well as trading-based measures (medium of trades, number of trades, turnover, volatility, drawdowns) both within a given fund, and across all funds. Considering performance as a proxy for sophistication, we will use quantiles of the distribution of past realized investor alphas to assign investors to the sophisticated and unsophisticated classes. We will then employ GBMs to identify the set of features that play an important role in determining investor sophistication.[4] Using the trained algorithm, we will classify investors as sophisticated and unsophisticated for the next period and gauge the out-of-sample performance of our algorithm.

For the second part of our analysis we will utilize the trading data and investor classification scheme to analyze the effect of mutual fund splits on individual investors' behavior. Potential lines of inquiry include: (1) Whether funds which split experience change in flows; (2) Whether these changes are temporary, and (3) Whether these changes are driven by investors we classify as sophisticated or unsophisticated on the eve of the split, relative to funds which did not split. We will also track the holding duration and subsequent outflows by investors of both classes. Under the marketability explanation, for instance, we would expect to see an increase in unsophisticated investment in funds which split, relative to non-splitting funds. These investors may, however, be more myopic, returns-sensitive, and trade more frequently than the existing fund investor base. Hence, such investors may close out their positions prematurely in response to downward market moves. This may impact medium- and long-run fund performance by forcing fund managers to inefficiently liquidate their holdings in order to satisfy abnormal redemptions by unsophisticated investors (as in Coval and Stafford, 2007). The response of sophisticated investors to splits is, *ex ante*, less clear. On one hand, these investors may anticipate the behavior of unsophisticated investors and invest less in (or withdraw from) funds which split, relative to those which do not. On the other hand, the failure of sophisticated investors to do so would shed light on unintended distributional consequences of splits — consequences which existing studies relying upon market data are unable to uncover. Taken together, our analysis aims to shed light on the effects of splits on investors, funds, and firms, including potentially heterogeneous consequences for various classes of investors.

[3] Another important contribution of our proposed methodology is that it can be readily generalized to other applications utilizing detailed investor-level data. For example, using the order level data in Boehmer, Li, and Saar (2018), we can generate features such as order size, frequency, number of stocks for each high-frequency trading firm and then use that to predict outcomes such as profitability. Unlike their principal-components-based analysis, our methodology has the added advantage in that we can comment on which of the observable features of trading firms and their strategies seem to be correlated with their profits.

[4] Our approach is flexible in that it allows for dynamics in (1) an investor being classified as sophisticated or not, accounting for potential learning through trading, and (2) which features matter at any given point in time.

References

Apley, D. W. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv preprint arXiv:1612.08468*.

Barber, B. M. and T. Odean (2000). Trading is Hazardous to your Wealth: The Common Stock Investment Performance of Individual Investors. *The Journal of Finance 55*(2), 773–806.

Barber, B. M. and T. Odean (2001). Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics 116*(1), 261–292.

Barber, B. M. and T. Odean (2002). Online Investors: Do the Slow Die First? *The Review of Financial Studies 15*(2), 455–488.

Barber, B. M., T. Odean, and L. Zheng (2005). Out Of Sight, Out Of Mind: The Effects of Expenses on Mutual Fund Flows. *The Journal of Business 78*(6), 2095–2120.

Biau, G., B. Cadre, and L. Rouvière (2018). Accelerated Gradient Boosting. *Machine Learning*, 1–22.

Boehmer, E., D. Li, and G. Saar (2018). The Competitive Landscape of High-Frequency Trading Firms. *The Review of Financial Studies 31*(6), 2227–2276.

Byun, J. and M. S. Rozeff (2003). Long-run Performance After Stock Splits: 1927 to 1996. *The Journal of Finance 58*(3), 1063–1085.

Chen, T. and C. Guestrin (2016). Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM.

Coval, J. and E. Stafford (2007). Asset Fire Sales (and Purchases) in Equity Markets. *Journal of Financial Economics 86*(2), 479–512.

Easley, D., M. Lopez de Prado, M. O'Hara, and Z. Zhang (2019). Microstructure in the Machine Age. *Working Paper*.

Easley, D., M. O'Hara, and G. Saar (2001). How Stock Splits Affect Trading: A Microstructure Approach. *Journal of Financial and Quantitative Analysis 36*(1), 25–51.

Fernando, C. S., S. Krishnamurthy, and P. A. Spindt (1999). Is Share Price Related To Marketability? Evidence From Mutual Fund Share Splits. *Financial Management*, 54–67.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189–1232.

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis 38*(4), 367–378.

Hutchinson, R. A., L.-P. Liu, and T. G. Dietterich (2011). Incorporating Boosted Regression Trees into Ecological Latent Variable Models. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Johnson, R. and T. Zhang (2014). Learning Nonlinear Functions Using Regularized Greedy Forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*(5), 942–954.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, pp. 3146–3154.

Lakonishok, J. and B. Lev (1987). Stock Splits and Stock Dividends: Why, Who, and When. *The Journal of Finance 42*(4), 913–932.

Ohlson, J. A. and S. H. Penman (1985). Volatility Increases Subsequent To Stock Splits: An Empirical Aberration. *Journal of Financial Economics 14*(2), 251–266.

Pittman, S. J. and K. A. Brown (2011). Multi-scale Approach for Predicting Fish Species Distributions Across Coral Reef Seascapes. *PLoS One 6*(5), e20583.

Rashmi, K. V. and R. Gilad-Bachrach (2015). DART: Dropouts meet Multiple Additive Regression Trees. In *AISTATS*, pp. 489–497.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). Why Should I Trust You?: Explaining the Predictions Of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM.

Rozeff, M. S. (1998). Stock Splits: Evidence from Mutual Funds. *The Journal of Finance 53*(1), 335–349.

Schultz, P. (2000). Stock Splits, Tick Size, and Sponsorship. *The Journal of Finance 55*(1), 429–450.

Weld, W. C., R. Michaely, R. H. Thaler, and S. Benartzi (2009). The Nominal Share Price Puzzle. *Journal of Economic Perspectives 23*(2), 121–42.