2024 ABFER

A Tale of Two Zoos: Machine Learning Insights on *15 million* Retail Investors

Pulak Ghosh (Indian Institute of Management Bangalore)
Huahao Lu (PBCSF, Tsinghua)
Hong Zhang (SMU)
Jian Zhang (The University of Hong Kong)

Motivation: A Tale of Two Zoos

- Factor Zoo—the "multidimensional challenge" (Cochrane 2011)
  - McLean and Pontiff (2015): 97 anomalies
  - Harvey, Liu, and Zhu (2016), 296
  - Hou, Xue, and Zhang (2018), 452
  - Jensen, Kelly, and Pedersen (2022): Global, 406 anomalies
- Bias Zoo—the "lack-of-discipline" concern (Fama, 1998)
  - Barber and Odean (2013); Hirshleifer (2015), and Barberis (2018) provide literature reviews
  - How to consolidate biases: Choi and Robertson (2020); Liu, Peng, Xiong, and Xiong, (2022)

Our Question

- Retail investors are in both zoos!
  - Balasubramaniam, Campbell, Ramadorai, and Ranish (2023) find strong investor clienteles for stock characteristics.
  - Other studies (e.g., Scandinavian accounts) document various biases.
- But which factors are the most important to investor welfare?
- To answer these questions, we need
  - A big data on retail investors
  - A powerful tool to digest retail investors' returns

Our Approach

- A **very big data** on retail investors
  - **15.4 million** valid retail investor accounts in India
- We employ a list of **ML tools** to **predict** retail investors' returns
  - Traditional linear (OLS) model
  - LASSO, Ridge, and Random Forest
  - Two *Neural Networks*
    - *Feedforward NN*
    - *Residual Neural Network (ResNN)*

Main Findings

- Neural Networks (esp. *ResNN*) outperform
  - They uniquely predict both good and bad out-of-sample performance.
  - Other models cannot predict good.
- Leading factors:
  - (Under)diversification, portfolio turnover, and momentum for overall retail returns
  - Turnover, the disposition effect, and diversification for the returns of newly initiated trading
- Behavioral biases > holding-weighted firm characteristics.

Road Map

- Data and variables
- Empirical Methods of using ML Models
- Empirical Analysis
  - Predicting Power of Models
  - Variable Gradient Analysis
  - Two Sources of Returns (holding vs. Trading)
  - Additional Analyses and Robustness Checks
- Conclusions

1. Data and Variables

- The National Stock Exchange of India(NSE): 2012-2020
  - Over **19 million** retail accounts, 7th largest worldwide.
  - We identify **15.4 million** valid retail investor accounts
  - Over **1.523 billion** investor-month return observations
- All listed stocks on NSE.
  - Prowess Database (similar to CRSP in the US) maintained by the Centre for Monitoring Indian Economy (CMIE).
  - Our main analysis **excludes 30%** of small stocks due to difficult-to-trade (Liu et al. 2019) and bias to machine learning models (Avramov, Cheng, and Metzker 2019 and Cong et al., 2020).
- We construct **23 holding-weighted stock characteristics** and **13 behavioral biases**/investor characteristics
  - DeMiguel, et al 2023: 17 mutual-fund characteristics
  - Kaniel, et al 2023: 46 stock characteristics and 13 fund and fund-family characteristics.

2. Predicting Models (1)

- Traditional **OLS**.
- **Lasso** and **Ridge**: introducing penalties for the magnitude of linear model parameters.

$$\min_{\beta \in R^p}\{\frac{1}{N}||y - X\beta||_2^2 + \lambda||\beta||_1 + \gamma||\beta||_2^2\},$$

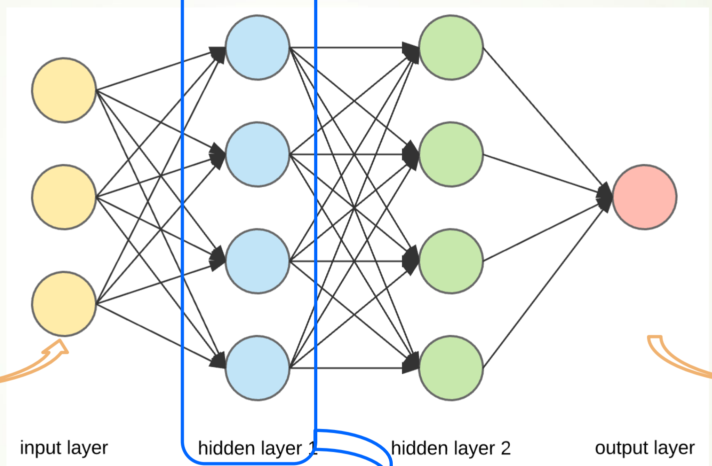
where β is the model parameters, Lasso: γ=0; Ridge: λ=0.

- **Random Forest**: constructing a multitude of decision trees during the training phase. It employ an ensemble strategy by averaging multiple deep decision trees, each trained on different segments of the same training set.

Predicting Models (2)

- **Feedforward Neural Network**:
  - A multi-layer perceptron network consists of an input layer, an output layer, and one or more hidden layers.

Initial inputs: firm characteristics & behavioral biases of each account



Output: return (ranks) of each account

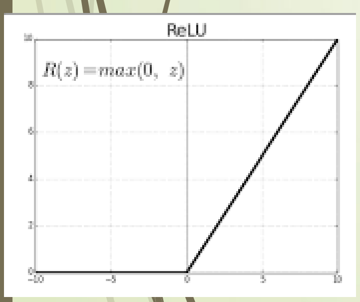
For the l-th layer: $\boxed{X^l} = g(W^{(l)T}\boxed{X^{(l-1)}} + b^{(l)})$

Output of the layer

input to the layer

$g(.)$: the non-linear activation function (ReLU)

$W^{(l)}$ and $b^{(l)}$: learnable parameters (weight + biases)



$$Relu(x) = \begin{cases} x, if\ x \geq 0 \\ 0, otherwise \end{cases}$$

Predicting Models (3)

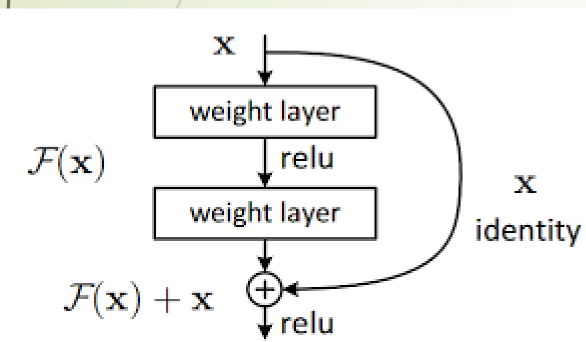
- **Residual Neural Network** (He et al., 2015):
  - Output for a layer = Residual + Input

$$\boxed{X^{l}} = \boxed{F\left(W^{(l)T}X^{(l-1)} + b^{(l)}\right)} + \boxed{X^{(l-1)}}$$

Output of the layer

"Residual": $F(.) = g(.)$- $X^{(l-1)}$

input to the layer



- Beneficial Features:
  - Each block tries to learn some **"new information" <u>(i.e., residuals)</u>** to **augment the data**, a simpler task to achieve with better information to learn from.
  - Each block has a **<u>shorter gradient path</u>**.
  - **<u>Modularity</u>** allows for deeper learning.

  - These features help address issues like **<u>overfitting</u>** and **<u>vanishing gradients</u>**.
  - Allows the algorithm to pay more attention to **economically important inputs** (e.g., biases)

Predicting Models (4):
Using models to predict retail performance

- Objective: use models to predict **total returns** of investors.
  - Following the literature (e.g., Kaniel et al., 2023), we train a model on 2/3 of the data and use it to predict returns on the remaining 1/3 subset. Hence, our tests are **out-of-sample**.
- Each model categorizes retail investors into five quintiles according to predicted returns.
  - The ***High*** and ***Low*** quintiles comprise the top and bottom 20% of predicted **winners** and **losers** among investors
  - We then calculate the **out-of-sample value-weighted returns** of the high/low groups of investors in the predicting period.
- The out of sample **High-minus-Low** returns indicate the predicting power of a model
  - We also use the locally estimated three-factor and four-factor models to adjust these returns.

3. Empirical Analysis

- Predicting Power of Models
- Variable Gradient Analysis
- Two Sources of Returns (holding vs. Trading)
- Additional Analyses and Robustness Checks

3.1 Out-of-sample Returns

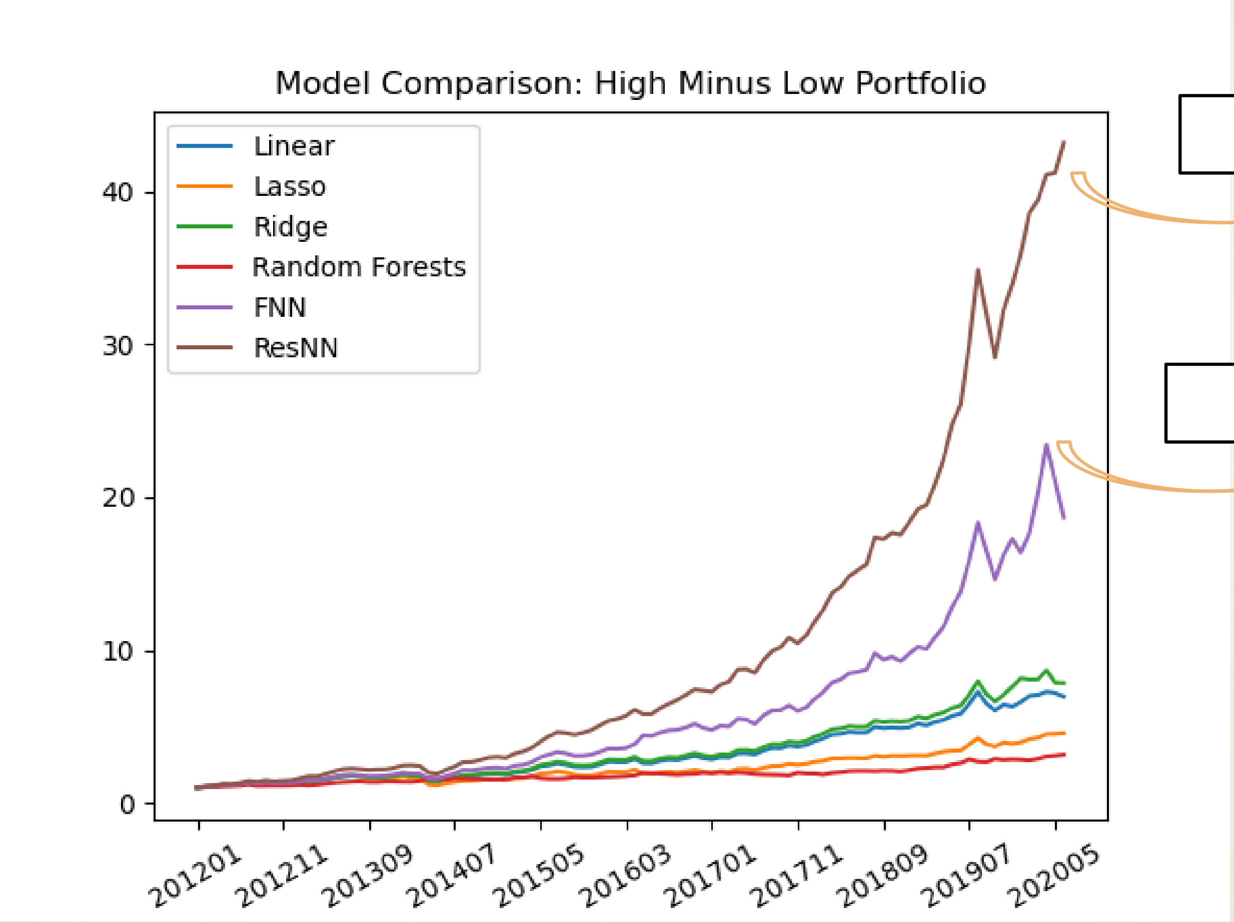
	(1)	(4)	(7)	(8)	(9)
	LOW	HIGH	High-minus-Low		
	Mean	Mean	Mean	FF-3	Carhart-4
Linear	-0.018*	0.007	0.025***	0.024***	0.022***
	(-1.77)	(1.37)	(3.34)	(3.14)	(2.95)
Lasso	-0.008	0.008	0.016**	0.014*	0.013*
	(-0.75)	(1.56)	(2.06)	(1.80)	(1.68)
Ridge	-0.018*	0.007	0.025***	0.024***	0.022***
	(-1.76)	(1.36)	(3.32)	(3.10)	(2.91)
Random Forest	-0.015	0.011	0.026	0.023	0.019
	(-1.24)	(1.08)	(1.47)	(1.59)	(1.63)
FNN	-0.025*	0.015***	0.040***	0.033***	0.031***
	(-1.74)	(2.66)	(3.38)	(3.24)	(3.11)
Residual Neural Network	-0.031**	0.012**	0.044***	0.042***	0.041***
	(-2.38)	(2.00)	(4.57)	(4.38)	(4.26)

ResNN is the best in predicting losers

NNs are the only model to predict winners!

Several Models can predict High-minus-Low. But NNs are the winners, particularly ResNN in economic magnitude.

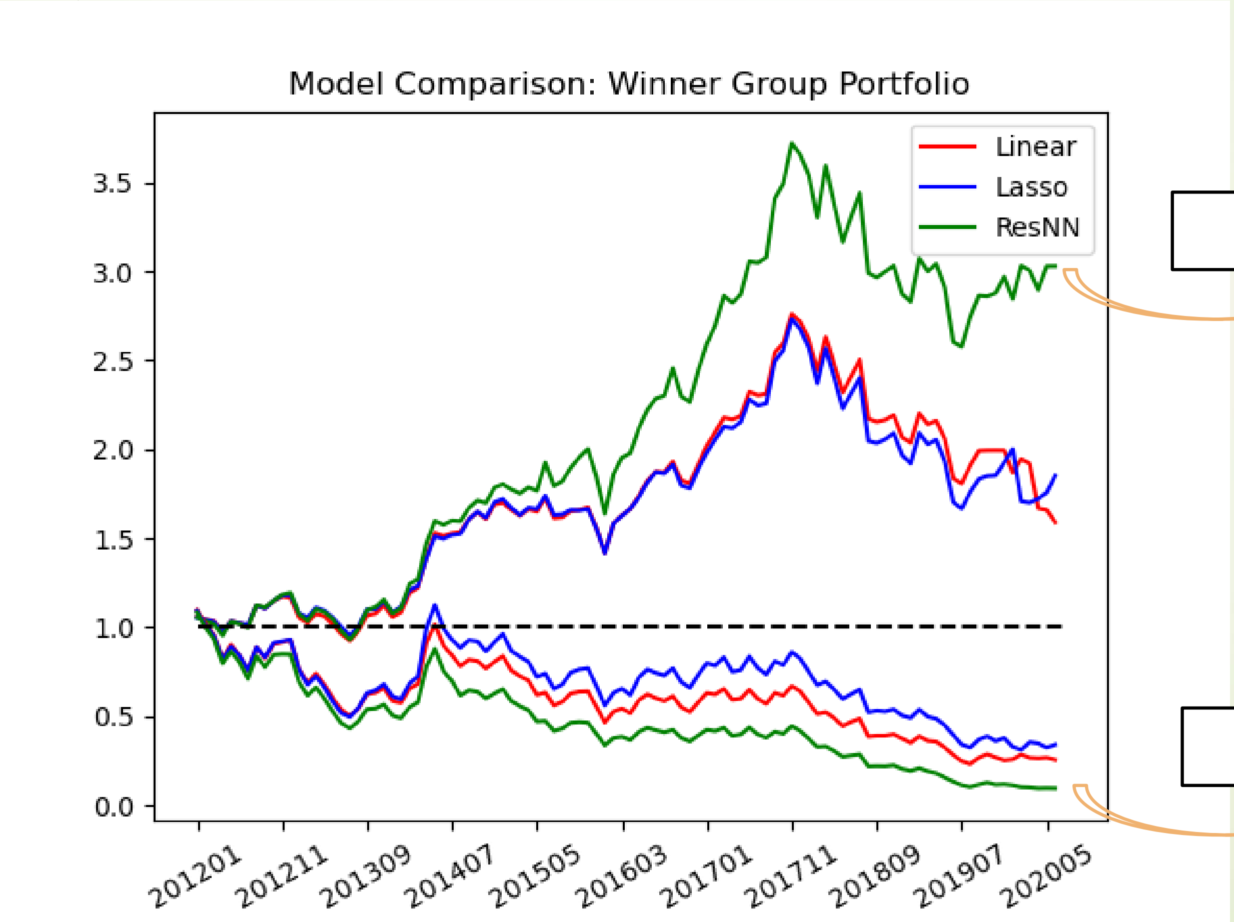
Figure 1: Cumulative Returns of High-minus-Low



Res_NN

FNN

Figure IN1: Zooming Into Cumulative Returns of High vs. Low



Res_NN

Res_NN

3.2 Which factors contribute more?

We use the traditional ***FNN*** to demonstrate the standalone predicting power of behavioral biases or firm characteristics.

	(1)	(3)	(4)	(6)	(7)	(9)	
	LOW		HIGH		H-L		
	Mean	Carhart-4	Mean	Carhart-4	Mean	Carhart-4	
FNN: Stock Char	-0.003	0.001	0.006	0.008	0.009	0.007	Stock Char alone cannot predict
	(-0.23)	(0.09)	(1.01)	(1.32)	(0.99)	(0.72)	
FNN: Behavioral	-0.025**	-0.022**	0.008	0.010*	0.033***	0.032***	Biases alone can predict
	(-2.62)	(-2.25)	(1.46)	(1.80)	(5.63)	(5.29)	
FNN: Stock Chars + Behavioral	-0.025*	-0.015	0.015***	0.017***	0.040***	0.031***	Better results when jointly used
	(-1.74)	(-1.43)	(2.66)	(3.08)	(3.38)	(3.11)	
ResNN: Stock Chars +Behavioral	-0.031**	-0.026**	0.012**	0.014**	0.044***	0.041***	Best results when Res NN is used
	(-2.38)	(-2.00)	(2.00)	(2.33)	(4.57)	(4.26)	

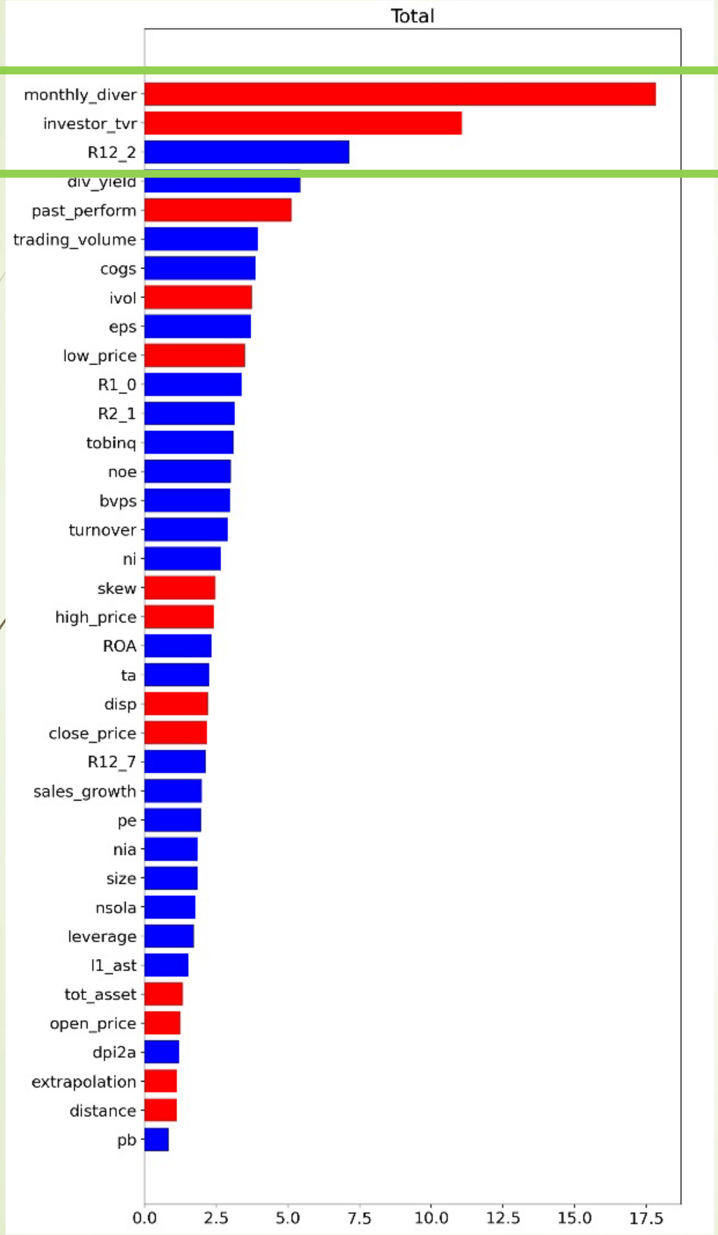
3.2 Variable Gradient Analysis:
The contribution of Individual Factors

- Sadhwani et al. (2020) and Horel and Giesecke (2020):

$$Importance(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(\frac{\partial R_{i,t+1}^{pred}}{\partial x_{i,t}}\right)^2,$$

where T represents the number of periods in the data, and N_t denotes the total number of investors in the t-th period.

Top Factors

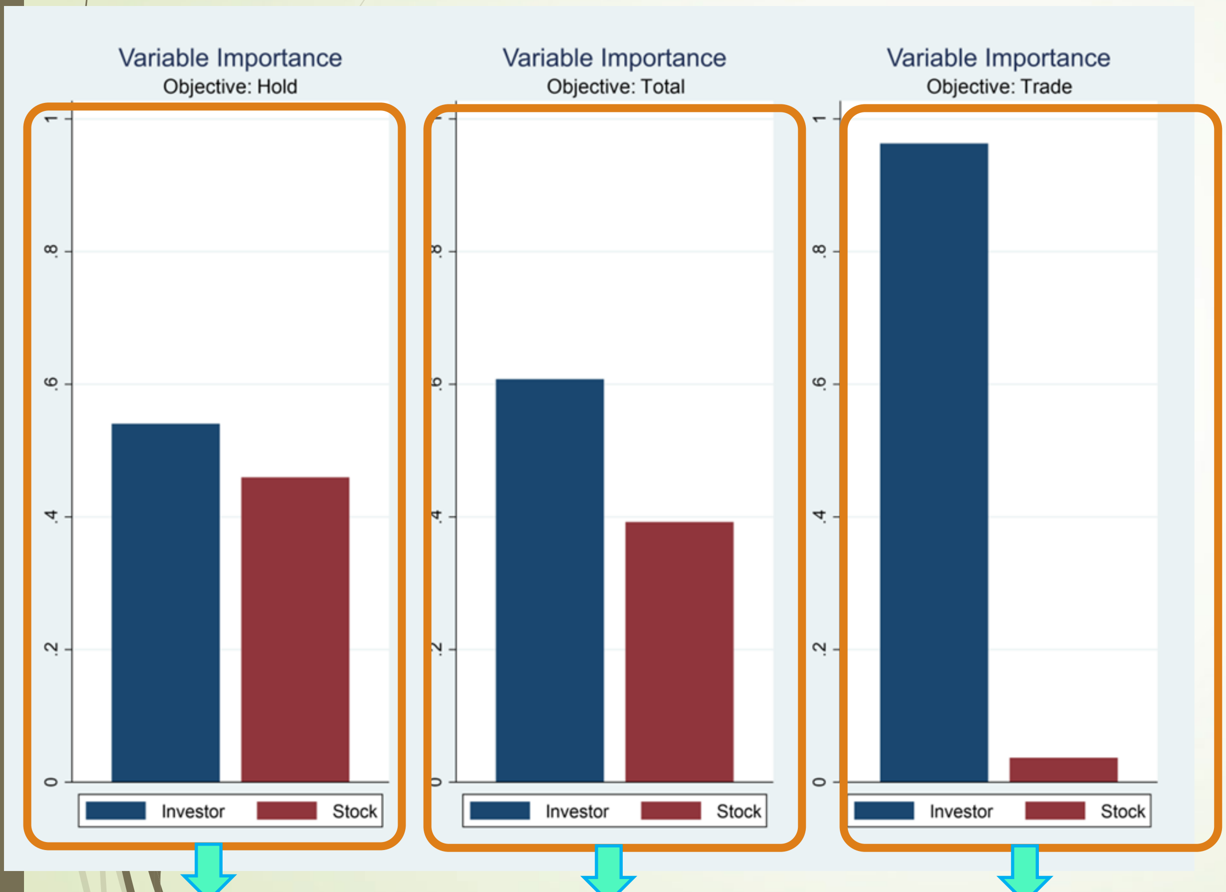


- We use different colors to highlight behavioral (red) vs. firm characteristics (blue).
- The Top 3 factors are:
  - (Under)diversification
  - Portfolio turnover
  - Momentum

3.3 Two Sources of Returns

- Investors' total month returns have two distinct sources:
  - Holding the existing portfolio (**holding based returns**)
  - New trades initiated during the month (i.e., **trading returns**).
- The two sources may be subject to different factors:
  - Selling and buying decision may be triggered by preference (e.g., the disposition effect or lottery preference) and new information (e.g., salience theory)
  - Holdings could allow firm char to play more role.

The Relative Importance of bias (blue bar) vs. firm Char (Red bar)



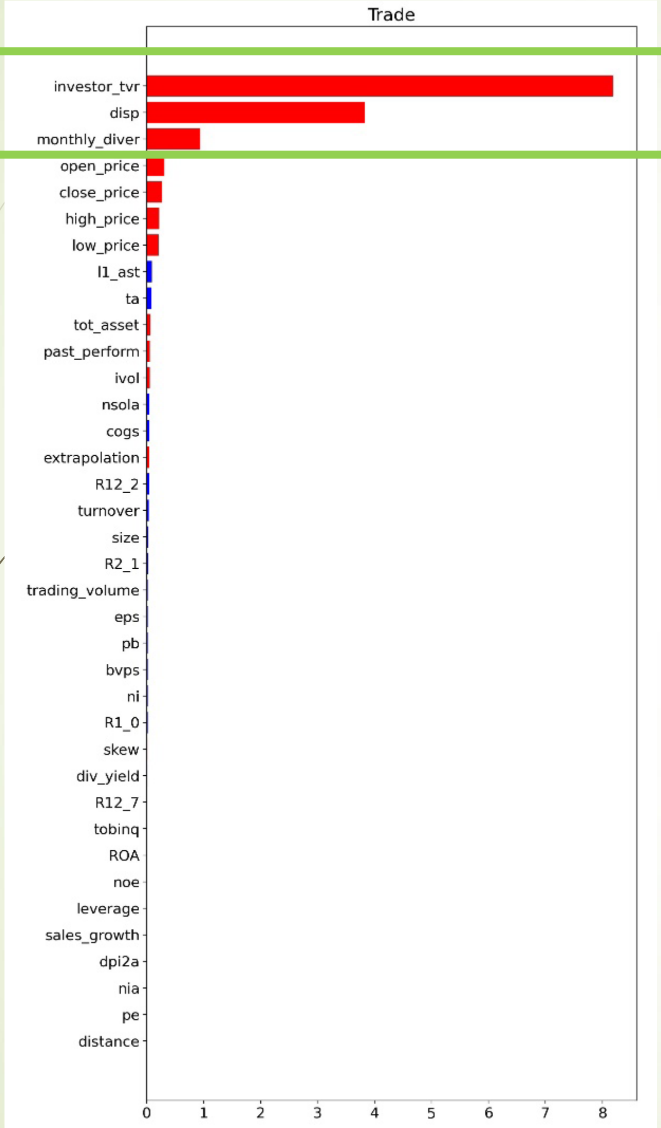
Training Goal: **Holding-based returns**

Training Goal: **Total returns**

Training Goal: **Trading returns**

- The joint explanatory power of behavioral increases in predicting: holding returns ➔ total returns ➔ trading returns.

What Drive New Trades



- Behavioral factors dominates the contributions to trading returns.
- The Top 3 factors are:
  - Portfolio turnover
  - The Disposition Effect
  - (Under)diversification
- All these factors contribute to bad total performance

Additional Analyses

- Model comparisons
- Robustness checks on removing small stocks (20%, 40%)
  - Robust
- Can NN trained on trading-returns or holding-returns also be used to predict total returns?
  - The answer is yes.
  - Hence, factors driving a particular element of return are also important for investor overall welfare.

Model comparison

	(4)	(5)	(6)
	High Minus Low (Excluding 30% Small Stocks)		
	Mean	FF-3	Carhart-4
FNN - Linear	0.015**	0.015***	0.016***
	(2.37)	(2.77)	(2.84)
FNN - Lasso	0.024***	0.019***	0.017***
	(2.95)	(3.11)	(3.17)
FNN - Ridge	0.015**	0.009**	0.010**
	(2.37)	(2.50)	(2.41)
FNN - Random Forest	0.014**	0.010**	0.013***
	(2.33)	(2.48)	(2.85)
ResNN - FNN	0.004*	0.009**	0.010***
	(1.79)	(2.26)	(2.83)

FNN outperforms other models in predicting the H-L returns

Between the two NN algorithms, ResNN outperforms.

Removing 20% or 40% small stocks does not change our reults

Conclusions

- We use various machine learning models to understand how behavioral heuristics and stock characteristics affect retail investors' investment returns.
- We find that Neural Networks (esp ResNN) uniquely predict both good and bad out-of-sample performance.
- Behavioral biases > holding-weighted firm characteristics. We also identify leading factors.
- Our analyses shed light on a unified and parsimonious framework to understand retail investors' investments and returns.