Presentation on paper “Empirical Asset Pricing Via Machine Learning”

Gu, Kelly, Xiu

U. Of Chicago, AQR CM

Agenda

- Motivation
- What is ML? Why apply ML? Which Methods are Used?
- Methodology
- Empirical Study of US Equities:
  - which indicators matter?
  - performance of models at stock level and aggregate
- Simulation (Appendix A)

Motivation

Provides comparative overview of ML methods applied to predicting equity returns, single stocks and aggregate portfolios, in the cross section and time series

Accuracy of ML methods may be evaluated by high out-of-sample predictive R^2 (and $Roos^2$) that are robust across some ML specifications

Identification of the better informative predictor variables for each method

Performance of methods

What is ML?

The definition of marchine learning is often context specific. The article uses this term to describe:

(i) A diverse collection of high-dimensional models for prediction

(ii) So-called “regularization” methods for model selection and mitigation of overfit

(iii) Efficient algorithms for serching among a vast number of potential model specifications

Why Apply ML to Asset Pricing?

Describe and understand differences in expected returns across assets

$Focuses\ on\ dynamics\ of\ singe\ stocks\ and\ the\ aggregate\ market\ equity\ risk\ premium$

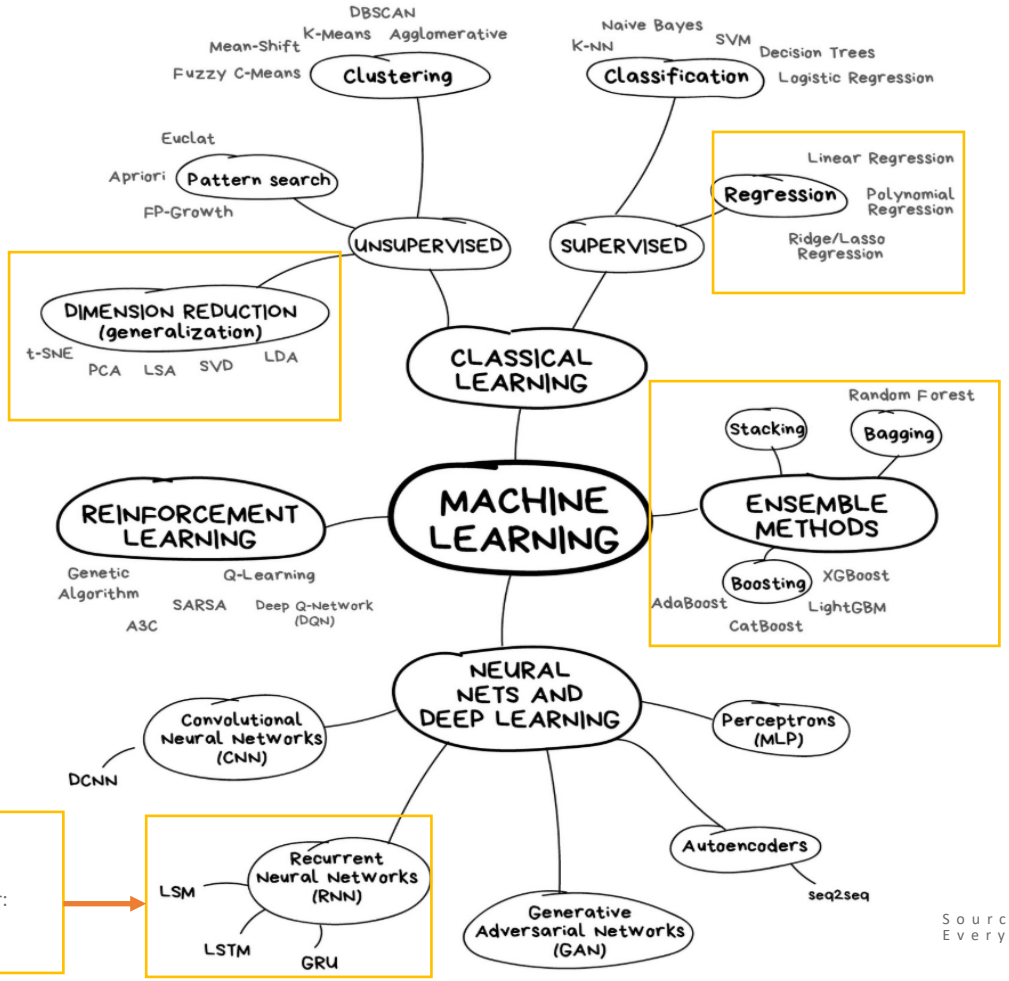
$Too\ many\ candidates\ (factor\ zoo)\ and, in\ many\ circunstances, not\ enough\ data.$
Dimensionality problem. Often highly correlated predictors

Many potential functional forms. Interaction among predictors

How ML can help?

- Suit of dissimilar methods generates a wide net in specification search
- May help deal with complex nonlinear associations
- May help with issues such as dimensionality reduction through parameter penalization; focus on variance when higher bias not the most important aspect

Which methods are used?



Feed-Forward NN
Up to 5 hidden layer:
NN1 thru NN5

Source: ML for Everyone

Teaser: including many variables impacts model predictability

30,000 individual stocks over 60 years (1957 to 2016). Predictor set with 94 characteristics, 74 industry sector dummies, 8 aggregate variables . Interactions total 900+ signals.
Panel regression of individual stock returns onto characteristics

Method	Tuning Hyperparameter	Performance (out-of-sample R^2 per month)
OLS panel Benchmark (size,BTM, Momentum)	No	0.16%
OLS panel 900+ predictors	No	negative
Elastic Net 900+ predictors	yes	0.09%
PCR/PLS 900+ predictors	yes	0.18%/0.28%
Trees 900+ predictors	yes	0.27%
NN 900+ predictors	yes	>> 0.30%

Teaser: Aggregate portfolio predictability improves with ML

Build bottom-up portfolio-level forecasts from stock-level forecasts
Bottom-up forecast of the S&P

Method	Tuning Hyperparameter	Performance (out-of-sample R^2 per month)
OLS panel Benchmark (size,BTM, Momentum)	No	…
OLS panel 900+ predictors	No	-0.11%
Generalized Linear	yes	0.86%
Trees 900+ predictors	yes	>1.39%
NN 900+ predictors	yes	>> 1.39%...to 1.8%

"More pronounced predictive power at the portfolio level (aggregate)vs the stock level is driven; individual stock returns behave erratically for smallest & least liquid stocks. Aggregating into portfolio averages out much of the unpredictive stock-level noise."

Methodology

First: choose statistical model

Second: Objective function for estimating model parameters
All of estimates share objective of minimizing mean squared prediction error (MSE)
Regularization is introduced through variations of the MSE objective

Design disjoint sub-samples:

- Training: used to estimate de model subject to a specific set of tuning parameter values
- Validation : used to change/tuning the hyperparameters. Construct forecasts for data points. Calculate the objective function based on forecast errors from validation sample. Evaluate using lowest MSE. Simulate out-of-sample
- Testing: re-train models in combined window (validation + test)
  evaluate a method´s predictive performance with best/optimal hyperparameters

Methodology: Performance Evaluation

To assess predictive performance for individual excess stock return forecasts, we calculate the out-of-sample R^2 as

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - \widehat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}}, \tag{19}$$

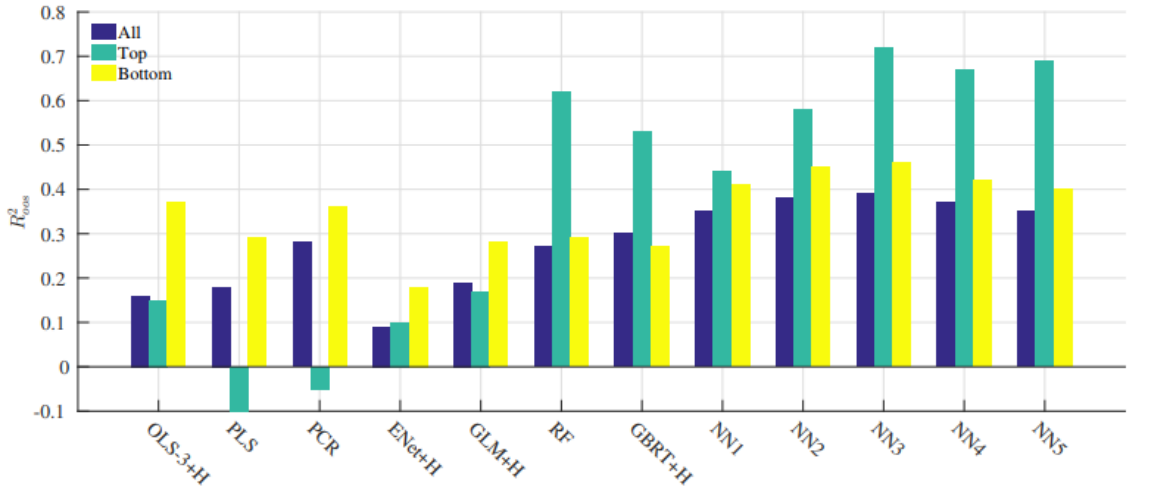
where $\mathcal{T}_3$ indicates that fits are only assessed on the testing subsample, whose data never enter into model estimation or tuning. The R^2_{oos} pools prediction errors across firms and over time into a grand panel-level assessment of each model.

A subtle but important aspect of our R^2 metric is that the denominator is the sum of squared excess returns *without demeaning*. In many out-of-sample forecasting applications, predictions are compared against historical mean returns. While this approach is sensible for the aggregate index or long-short portfolios, for example, it is flawed when it comes to analyzing individual stock returns. Predicting future excess stock returns with historical averages typically *underperforms* a naive forecast of zero by a large margin. That is, the historical mean stock return is so noisy that it artificially lowers the bar for "good" forecasting performance. We avoid this pitfall by benchmarking our R^2 against a forecast value of zero. To give an indication of the importance of this choice, when we benchmark model predictions against historical mean stock returns, the out-of-sample monthly R^2 of all methods rises by roughly three percentage points.

Empirical Study of US Equities

Table 1: Monthly Out-of-sample Stock-level Prediction Performance (Percentage R^2_{oos})

	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	-4.60	0.16	0.18	0.28	0.09	0.19	0.27	0.30	0.35	0.38	0.39	0.37	0.35
Top 1000	-14.21	0.15	-0.10	-0.05	0.10	0.17	0.62	0.53	0.44	0.58	0.72	0.67	0.69
Bottom 1000	-2.13	0.37	0.29	0.36	0.18	0.28	0.29	0.27	0.41	0.45	0.46	0.42	0.40



Note: In this table, we report monthly R^2_{oos} for the entire panel of stocks using OLS with all variables (OLS), OLS using only size, book-to-market, and momentum (OLS-3), PLS, PCR, elastic net (ENet), generalize linear model (GLM), random forest (RF), gradient boosted regression trees (GBRT), and neural networks with one to five layers (NN1–NN5). "+H" indicates the use of Huber loss instead of the l_2 loss. We also report these R^2_{oos} within subsamples that include only the top 1,000 stocks or bottom 1,000 stocks by market value. The lower panel provides a visual comparison of the R^2_{oos} statistics in the table (omitting OLS due to its large negative values).

Empirical Study of US Equities

Table 3: Comparison of Monthly Out-of-Sample Prediction using Diebold-Mariano Tests

	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
OLS+H	**3.81**	**3.82**	**3.85**	**3.81**	**3.83**	**3.91**	**3.94**	**3.96**	**3.96**	**3.98**	**3.97**	**3.96**
OLS-3+H		0.23	1.72	-0.80	0.63	1.55	1.93	**1.98**	**2.83**	**3.01**	**2.61**	**2.63**
PLS			1.58	-0.71	0.08	1.39	1.61	1.52	**2.29**	**2.43**	**2.18**	**2.15**
PCR				-1.51	-1.62	0.06	0.48	0.54	1.13	1.20	0.94	0.85
ENet+H					1.00	1.59	1.79	**2.09**	**2.02**	**2.19**	1.92	1.94
GLM+H						1.21	1.59	1.70	**2.55**	**2.76**	**2.44**	**2.33**
RF							0.66	0.66	1.12	1.30	0.94	0.90
GBRT+H								0.24	0.73	0.83	0.53	0.46
NN1									0.87	1.11	0.49	0.31
NN2										0.10	-1.09	-1.20
NN3											-1.03	-1.92
NN4												-0.47

Note: This table reports pairwise Diebold-Mariano test statistics comparing the out-of-sample stock-level prediction performance among thirteen models. Positive numbers indicate the column model outperforms the row model. Bold font indicates the difference is significant at 5% level or better.

NN are the only models that produce large and significant improvements over linear and generalized linear

Which Indicators Matter

Methods produce similar selection of the most informative stock-level indicators

1. First predictors are price trend such as stock momentum, industry momentum, short-term reversal

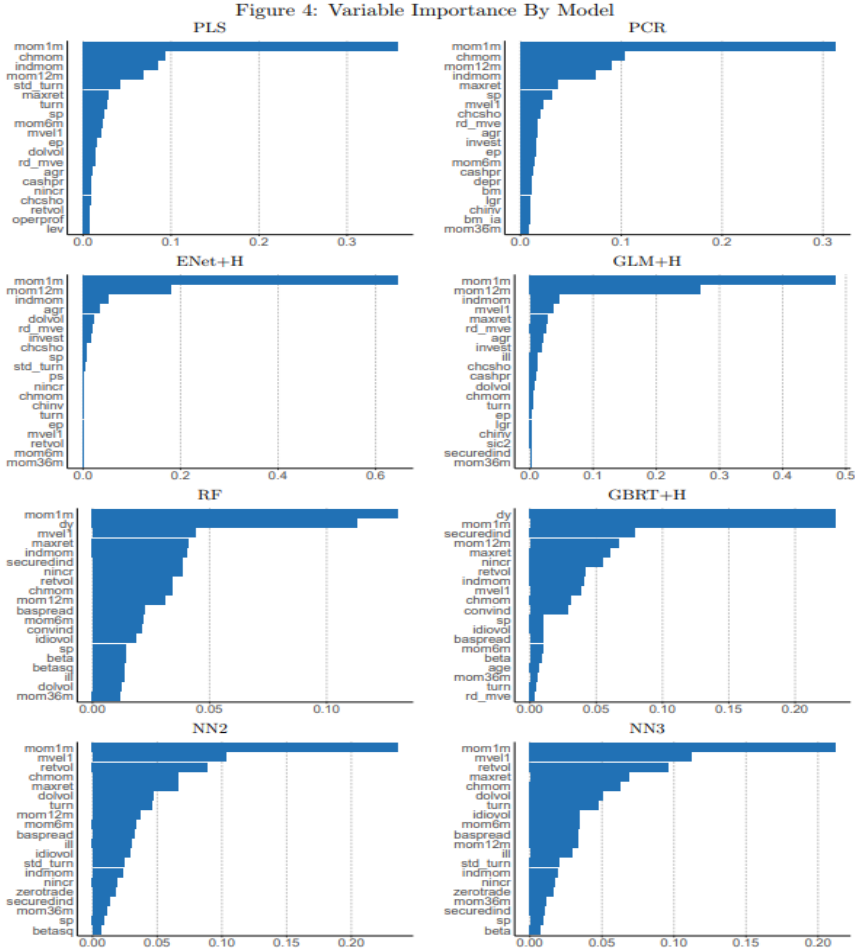
2. Next liquidity indicators such as turnover, dollar volume, bid/ask spread, etc

3. Next risk measures such as beta, volatility, etc

4. Last group includes valuation ratios and fundamental signals

Which Indicators Matter

Figure 4: Variable Importance By Model



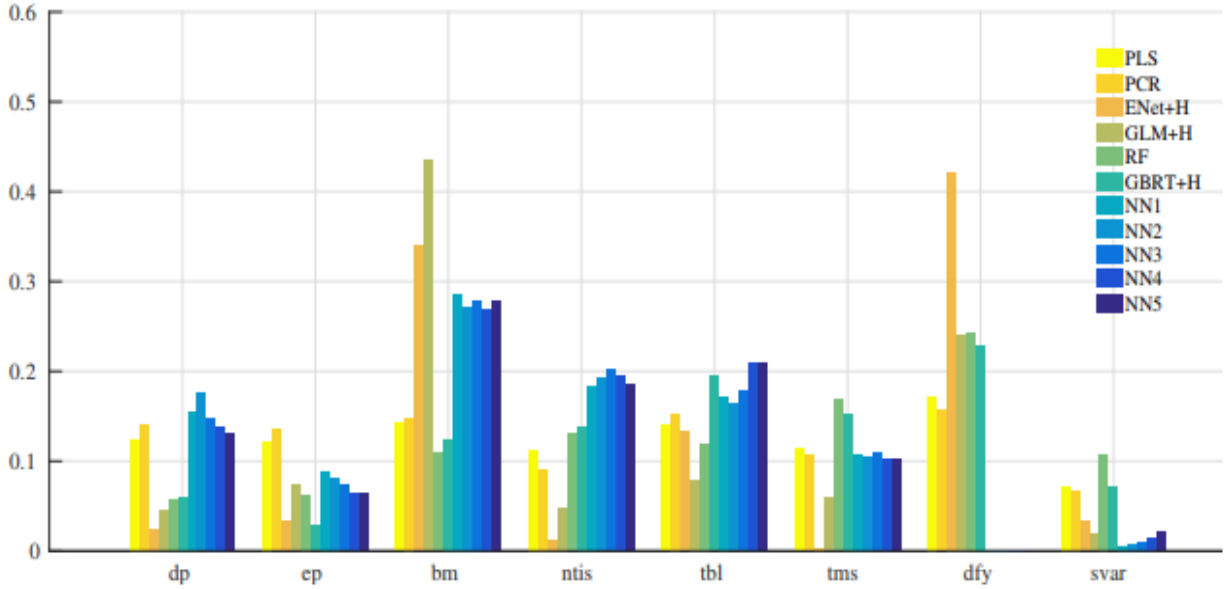
Note: Variable importance for the top 20 most influential variables in each model. Variable importance is an average over all training samples. Variable importances within each model are normalized to sum to one.

Indicators importance magnitude for penalized linear models and dimension reduction models are skewed toward momentum and reversal

Trees and NN are more democratic (see bars), drawing predictive information from a broader set of characteristics, but still skewed to the 3 “families” (momentum, liquidity, risk)

Which Indicators Matter, for aggregates/macro

All methods agree on aggregate BTM; while Linear favor aggregate fundamental, Trees and NN also favor treasury rates and term spread. We know that expectations improve this analysis



Note: Variable importance for eight macroeconomic variables in each model. Variable importance is an average over all training samples. Variable importances within each model are normalized to sum to one. The lower panel provides a complementary visual comparison of macroeconomic variable importances.

Portfolio Level Performance

Aggregating individual stock into portfolios. Forecasts for aggregate portfolios including S&P, six size & value, six size & investment, six size & profit., six size & momentum.

Table 5: Monthly Portfolio-level Out-of-Sample Predictive R^2

	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
S&P 500	-0.11	-0.86	-2.62	-0.38	0.86	1.39	1.13	0.84	0.96	1.80	1.46	1.60
Big Growth	0.41	0.75	-0.77	-1.55	0.73	0.99	0.80	0.70	0.32	1.67	1.42	1.40
Big Value	-1.05	-1.88	-3.14	-0.03	0.70	1.41	1.04	0.78	1.20	1.57	1.17	1.42
Big Neutral	0.12	-0.81	-2.39	-0.46	0.41	1.05	1.03	1.33	0.78	1.81	1.93	1.93
Small Growth	0.35	1.54	0.72	-0.03	0.95	0.54	0.62	1.68	1.26	1.48	1.53	1.44
Small Value	-0.06	0.40	-0.12	-0.57	0.02	0.71	0.90	0.00	0.47	0.46	0.41	0.53
Small Neutral	-0.01	0.78	-0.10	-0.25	0.36	0.41	0.38	0.58	0.55	0.68	0.62	0.72
Big Conservative	-0.24	-0.17	-1.97	0.19	0.69	0.96	0.78	1.08	0.67	1.68	1.46	1.56
Big Aggressive	-0.12	-0.77	-2.00	-0.91	0.68	1.83	1.45	1.14	1.65	1.87	1.55	1.69
Big Neutral	-0.36	-1.65	-3.20	-0.11	0.76	0.99	0.73	0.54	0.62	1.62	1.44	1.60
Small Conservative	0.02	0.75	0.48	-0.46	0.55	0.59	0.60	0.94	0.91	0.93	0.99	0.88
Small Aggressive	0.14	0.97	0.06	-0.54	0.19	0.86	1.04	0.25	0.66	0.75	0.67	0.79
Small Neutral	-0.04	0.53	-0.17	0.08	0.45	0.23	0.20	0.73	0.60	0.81	0.73	0.80
Big Robust	-0.58	-0.22	-2.89	-0.27	1.54	1.41	0.70	0.60	0.84	1.14	1.05	1.21
Big Weak	-0.24	-1.47	-1.95	-0.40	-0.26	0.67	0.83	0.24	0.60	1.21	0.95	1.07
Big Neutral	-0.08	-1.02	-2.77	-0.21	0.10	1.46	1.44	0.95	1.00	1.78	1.70	1.73
Small Robust	-0.77	0.77	0.18	-0.32	0.41	0.27	-0.06	-0.06	-0.02	0.06	0.13	0.15
Small Weak	0.02	0.32	-0.28	-0.25	0.17	0.90	1.31	0.84	0.85	1.09	0.96	1.08
Small Neutral	0.22	1.05	0.09	0.03	0.48	0.76	0.97	1.08	1.04	1.19	1.12	1.18
Big Up	-1.53	-2.54	-3.93	-0.21	0.40	1.12	0.68	0.46	0.85	1.28	0.99	1.05
Big Down	-0.10	-1.20	-2.05	-0.26	0.36	1.09	0.77	0.48	0.89	1.34	1.17	1.36
Big Medium	0.24	1.38	0.57	0.01	1.32	1.56	1.37	1.60	1.76	2.28	1.83	2.01
Small Up	-0.79	0.42	-0.36	-0.33	-0.33	0.31	0.40	0.23	0.60	0.67	0.55	0.61
Small Down	0.40	1.16	0.47	-0.46	0.62	0.93	1.20	0.80	0.97	0.97	0.97	0.96
Small Medium	-0.29	0.03	-0.61	-0.56	-0.20	0.11	0.18	0.05	0.29	0.41	0.30	0.45

Non-linear methods produce better out-of-sample predictions for market and aggregate portfolios

Portfolio Level Performance

Forecasts for aggregate portfolios including S&P, six size & value, six size & investment, six size & profit., six size & momentum.

Table 6: Implied Sharpe Ratio Improvements

	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
S&P 500	-	-	-	-	0.11	0.17	0.14	0.11	0.12	0.21	0.17	0.19
Big Growth	0.05	0.09	-	-	0.08	0.11	0.09	0.08	0.04	0.18	0.15	0.15
Big Value	-	-	-	-	0.10	0.19	0.15	0.12	0.17	0.21	0.16	0.19
Big Neutral	0.02	-	-	-	0.06	0.13	0.13	0.17	0.10	0.22	0.23	0.23
Small Growth	0.03	0.14	0.07	-	0.09	0.05	0.06	0.15	0.11	0.13	0.13	0.13
Small Value	-	0.08	-	-	0.00	0.14	0.17	0.00	0.10	0.09	0.09	0.11
Small Neutral	-	0.08	-	-	0.04	0.04	0.04	0.06	0.06	0.07	0.07	0.08
Big Conservative	-	-	-	0.02	0.08	0.11	0.09	0.12	0.08	0.18	0.15	0.16
Big Aggressive	-	-	-	-	0.11	0.26	0.22	0.18	0.24	0.26	0.23	0.24
Big Neutral	-	-	-	-	0.10	0.12	0.09	0.07	0.08	0.19	0.17	0.19
Small Conservative	0.00	0.08	0.05	-	0.06	0.06	0.07	0.10	0.10	0.10	0.11	0.09
Small Aggressive	0.03	0.15	0.01	-	0.04	0.14	0.16	0.05	0.11	0.12	0.11	0.13
Small Neutral	-	0.05	-	0.01	0.04	0.02	0.02	0.07	0.06	0.08	0.07	0.08
Big Robust	-	-	-	-	0.17	0.16	0.08	0.07	0.10	0.13	0.12	0.14
Big Weak	-	-	-	-	-	0.12	0.14	0.05	0.11	0.19	0.16	0.17
Big Neutral	-	-	-	-	0.02	0.20	0.20	0.14	0.15	0.24	0.23	0.23
Small Robust	-	0.08	0.02	-	0.04	0.03	-	-	-	0.01	0.01	0.02
Small Weak	0.00	0.06	-	-	0.03	0.16	0.21	0.15	0.15	0.18	0.16	0.18
Small Neutral	0.03	0.12	0.01	0.00	0.06	0.09	0.11	0.12	0.12	0.13	0.12	0.13
Big Up	-	-	-	-	0.05	0.13	0.08	0.06	0.10	0.14	0.11	0.12
Big Down	-	-	-	-	0.06	0.15	0.11	0.07	0.13	0.18	0.16	0.18
Big Medium	0.05	0.23	0.11	0.00	0.22	0.25	0.23	0.26	0.28	0.34	0.29	0.31
Small Up	-	0.05	-	-	-	0.04	0.05	0.03	0.07	0.08	0.06	0.07
Small Down	0.07	0.17	0.08	-	0.10	0.15	0.18	0.13	0.15	0.15	0.15	0.15
Small Medium	-	0.00	-	-	-	0.01	0.02	0.01	0.03	0.05	0.04	0.05

Note: Improvement in annualized Sharpe ratio ($SR^* - SR$) implied by the full sample Sharpe ratio of each portfolio together with machine learning predictive R^2_{oos} from Table 5. Cases with negative R^2_{oos} imply a Sharpe ratio deterioration and are omitted.

Non-linear methods produce higher performance of aggregate portfolios

Simulation (Appendix A)

Monte Carlo simulation of 3 factor model in two versions : (a) indicators enter only linearly and additively; (b) indicators enter thru nonlinear & pairwise interactions

Table A.1: Comparison of Predictive R^2s for Machine Learning Algorithms in Simulations

Model	(a)				(b)			
Parameter	$P_c = 50$		$P_c = 100$		$P_c = 50$		$P_c = 100$	
R^2(%)	IS	OOS	IS	OOS	IS	OOS	IS	OOS
OLS	7.50	1.14	8.19	-1.35	3.44	-4.72	4.39	-7.75
OLS+H	7.44	1.25	8.08	-1.16	3.38	-4.56	4.27	-7.50
PCR	2.69	0.90	1.70	0.43	0.65	0.02	0.41	-0.01
PLS	6.24	3.48	6.19	2.82	1.02	-0.08	0.99	-0.17
Lasso	6.04	4.26	6.08	4.25	1.36	0.58	1.36	0.61
Lasso+H	5.96	4.27	6.00	4.26	1.27	0.59	1.27	0.62
Ridge	6.46	3.89	6.67	3.39	1.66	0.34	1.76	0.23
Ridge+H	6.36	3.90	6.54	3.40	1.58	0.35	1.67	0.25
ENet	6.04	4.26	6.08	4.25	1.35	0.58	1.35	0.61
ENet+H	5.96	4.27	6.00	4.26	1.27	0.59	1.27	0.62
GLM	5.91	4.11	5.94	4.08	3.38	1.22	3.31	1.17
GLM+H	5.84	4.13	5.88	4.10	3.32	1.27	3.23	1.22
RF	8.37	3.37	8.23	3.27	8.09	3.03	8.29	3.06
GBRT	7.09	3.35	7.04	3.28	6.50	2.76	6.41	2.82
GBRT+H	7.16	3.46	7.09	3.39	6.47	3.12	6.37	3.22
NN1	6.54	4.42	6.78	4.31	5.57	2.77	5.82	2.60
NN2	6.50	4.40	6.71	4.30	6.31	3.12	6.40	2.91
NN3	6.48	4.37	6.63	4.22	6.02	2.97	6.12	2.70
NN4	6.49	4.32	6.63	4.17	5.96	2.82	6.11	2.56
NN5	6.43	4.27	6.60	4.19	5.71	2.65	5.65	2.20
Oracle	6.22	5.52	6.22	5.52	5.86	5.40	5.86	5.40

Note: In this table, we report the average in-sample (IS) and out-of-sample (OOS) R^2s for models (a) and (b) using Ridge, Lasso, Elastic Net (ENet), generalized linear model with group lasso (GLM), random forest (RF), gradient boosted regression trees (GBRT), and five architectures of neural networks (NN1,...,NN5), respectively. "+H" indicates the use of Huber loss instead of the l_2 loss. "Oracle" stands for using the true covariates in a pooled-OLS regression. We fix $N = 200$, $T = 180$, and $P_x = 2$, comparing $P_c = 100$ with $P_c = 50$. The number of Monte Carlo repetitions is 100.

Model (a) Lasso & ENET deliver best oosR2. Huber Loss improves Lasso, Ridge, ENET. Other models tend to overfit

Model (b) RF, NNs perform better

OLS worst. PCR & PLS to not perform well