DNN, GBT, RF, Statistical Arbitrage on the S&P 500

Overview of the paper

RunPing Zhao 211232011
Haohan Zhao 211098172

Nanjing University

April 23, 2024

南京大学
NANJING UNIVERSITY

Statistical Arbitrage using Machine Learning

- In recent years, machine learning & deep learning have gained momentum, introducing a variety of renowned methods applicable across multiple fields, including finance.
- This paper implements and analyzes the effectiveness of ***Deep Neural Networks (DNN)***, ***Gradient Boosted Trees (GBT)***, ***Random Forests (RAF)***, and several ***ensembles*** of these methods in the context of statistical arbitrage. Each model is trained on lagged returns of all stocks in the S&P 500, after elimination of survivor bias.
- From 1992 to 2015, based on the probability forecast of a stock to outperform the general market, daily one-day-ahead trading signals are generated. Then long the highest k and short the lowest k. This is the basic of statistical arbitrage.

Background

- Statistical Arbitrage, a quantitative trading strategies, which encompasses strategies with the folowing features:
  - Trading signals are systematic, or rule-based.
  - The trading book is market-neutral (zero beta).
  - The mechanism for generating excess returns is statistical.

  Large numbers of securities, very short holding periods, and substantial computational, trading, and information technology (IT) infrastructure.
- An evolving gap between academic finance and financial industry.
  - Classical financial research is primarily focused on identifying capital market anomalies with high explanatory value, relying on linear models or (conditional) portfolio sorts, based on monthly data.
  - The latter are prone to deploy black-box methods on the short-term for the sake of profitability.

Brief Overview

This paper attempts to bridge this gap: it develops a short term *StatArb* strategy for the S&P 500 constituents.

- ***Deep Neural Networks (DNNs)*** – a type of highly parameterized neural network composed of multiple hidden layers, thus allowing for feature abstraction.
- ***Gradient Boosted Trees (GBTs)*** – combining many weak learners into one strong learner.
- ***Random Forests (RFs)*** – a large collection of de-correlated learners.
- ***Ensemble*** – equal weighted (ENS1), performance based (ENS2), rank based (ENS3).

This paper trains models with lagged returns of all stocks and forecast the probability for each stock to outperform the general market. The highest k stocks are bought and the lowest k stocks are sold.

Related Literature

- Huck (2009) develops a statistical arbitrage strategy based on ensembles of Elman neural networks and ELECTRE III.
- Takeuchi and Lee (2013) develop an enhanced momentum strategy on the U.S. CRSP stocks from 1965 until 2009.
- Moritz and Zimmermann (2014) deploy random forests on U.S. CRSP data from 1968 to 2012 to develop a trading strategy relying on ”deep conditional portfolio sorts”.
- Krauss (2016) provides a recent review of more than 90 statistical arbitrage pairs trading strategies.
- Atsalakis and Valavanis (2009) survey over 100 papers employing machine learning techniques for stock market forecasting.
- Sermpinis, Theofilatos, Karathanasopoulos, Georgopoulos, and Dunis (2013) provide further references in this respect.

Contribution

This study is dedicated to bridge the gap between academic and professional finance.

- Up to that time, this study is unique in deploying three state-of-the-art machine learning techniques and different ensembles on a large and liquid stock universe.
- It provides a holistic performance evaluation, following current standards in the financial literature.
- It focuses on a daily investment horizon instead of monthly frequencies, allowing for much more training data and for profitably exploiting short-term dependencies.

Data

For reasons about computational feasibility, market efficiency, and liquidity, S&P 500 stock universe is chosen. Data is prepared as follows:

1. Eliminate survivor bias following Krauss and Stübinger (2015).
2. Obtain all month end constituent lists for the S&P 500 from Thomson Reuters Datastream from December 1989 to September 2015, then consolidate these lists into one binary matrix, indicating whether the stock is a constituent of the index in the subsequent month or not.
3. For all stocks having ever been a constituent of the index, download the daily total return indices from January 1990 until October 2015.
4. Report average monthly summary statistics following Clegg and Krauss (2016)

Software

- Preprocessing and data handling are conducted in *R*.
- Using the packages *xts* and *TTR* for time series subsetting.
- Employ several routines in the package *PerformanceAnalytics* for performance evaluation.
- Deep neural networks, gradient-boosted trees, and random forests are implemented via *H2O*.
- Part of the communication between R and H2O is implemented with *Windows PowerShell*.
- We will use *Python* and *Pytorch* to implement all of our works.

Split Training and Trading Sets

- ”Study Period” is defined as a training-trading set, consisting of a 750 day training period (approximately three years), and a subsequent 250 day trading period (approximately one year).
- Split the entire data set from 1990 until 2015 in 23 of these study periods, so that the 250 day trading periods are non-overlapping.
- Let n_i denote the number of stocks in the S&P 500 at the end of the training period of study period $i \in \{1, \cdots, 23\}$, having full price information available. Typically, n_i is close to 500 for all study periods. Clearly, the stocks considered in each of these 23 batches are time-varying, depending on index constituency and full data availability.
- We will expand the time range to 1990-2023.

Feature Engineering

For each study period, the feature space (input) and the response variable (output) are as follows:

- *Input*: Denote $P^s = (P_t^s)_{t \in T}, s \in \{1, \cdots, n\}$ as the price series of stock s. Define simple return $R_{t,m}^s$ for each stock s over m periods as follow and get 31 features (losing 240 days):

$$R_{t,m}^s = \frac{P_t^s}{P_{t-m}^s} - 1, \quad m \in \{\{1, \cdots, 20\} \cap \{40, 60, \cdots, 240\}\} \quad (1)$$

- *Output*: A binary variable $Y_{t+1}^s \in \{0, 1\}$ for stock s, which equals 1 when its return $R_{t+1,1}^s > median\ return$.
  Purpose: Forecast a probability $\mathcal{P}_{t+1|t}^s$ for stock s to outperform the cross-sectional median in period $t + 1$.
- Training days: 500; Trading days: 250.
- Training sets: $510 \times 500 \times 32$; Trading sets: $250 \times 500 \times 31$.

Deep Neural Networks – Basic Conceptions

- Topology of a net: an input layer, one or more hidden layers, and an output layer.
- Input layer matches the feature space, having the same number of neurons as features.
- Output layer matches the output space, either a regression or a classification layer.
- Hidden layers are connected one-by-one, so as to implement feature abstraction.
- All layers are composed of neurons, which will output a weighted combination of n_l outputs of layer l, plus a bias b:

$$\alpha = \sum_{i=1}^{n_l} w_i x_i + b \tag{2}$$

Deep Neural Networks – Configurations

- ***Model Structure*** – I–H1–H2–H3–O with 31-31-10-5-2 neurons, 2746 parameters in total, so 93 training examples per parameter.
- ***Activation Function*** – MAXOUT:

$$f(\alpha_1, \alpha_2) = \max(\alpha_1, \alpha_2), \quad f: \mathbb{R}^2 \to \mathbb{R} \tag{3}$$

- ***Loss Function*** – Cross-Entropy: Let W, B denote the collection of weight matrices and biases, then:

$$\mathscr{L}(W, B|j) = -\sum_{y \in \mathscr{O}} \left(ln(o_y^{(j)}) t_y^{(j)} + ln(1 - o_y^{(j)})(1 - t_y^{(j)}) \right) \tag{4}$$

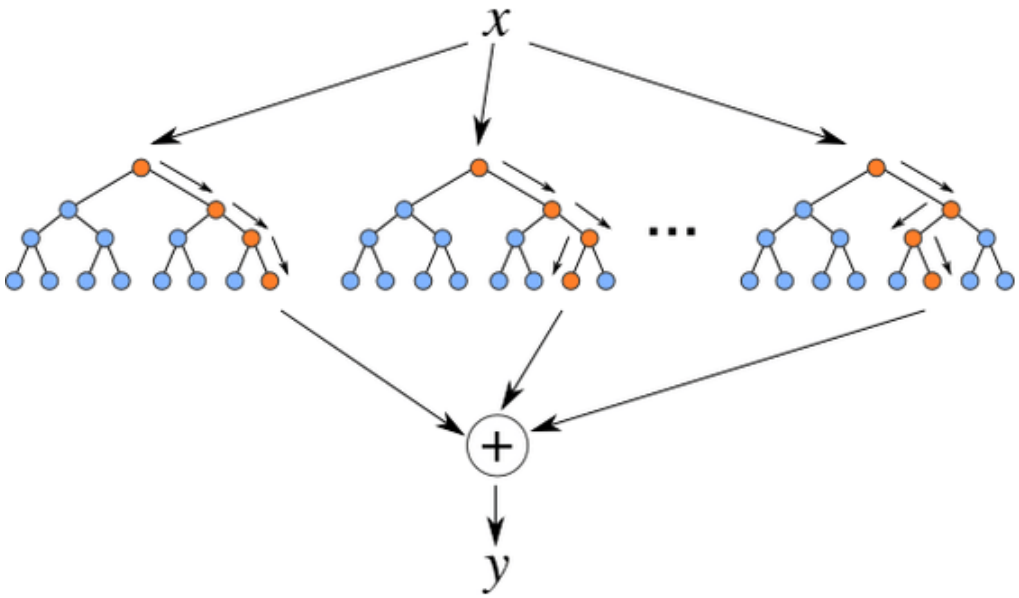
- ***Optimization*** – Stochastic Gradient Descent, ADADELTA[1].
- ***Regularization*** – *Dropout* with drop out rate 0.1 for input layer and 0.5 for others; *L1* with shrinkage rate $\lambda_{DNN} = 0.00001$.
- ***Epochs*** – 400, with an early stopping strategy.

[1] *H2O*'s optimizer, with its default setting

Gradient Boosted Trees

- ***Boosting***: An ensemble technique for "converting a weak learning algorithm into one that achieves arbitrarily high accuracy". The key idea behind boosting is to focus more on training instances that the predecessor model misclassified or had a higher error, thus giving more weight to those instances in the subsequent model.
  1. Initiate $f_0(x) = 0$. Set Loss Function $L(y, f(x))$.
  2. for $m = 1, 2, \cdots, M$:
     1. Compute negative gradient: $-g_m(x_i) = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$
     2. Train a regression tree $T_m(x, \Theta)$ with negative gradient $-g_m(x)$ as target.
     3. Let $f_m(x) = f_{m-1}(x) + \rho T_m(x, \Theta)$
  3. Get the final model after M iterations: $f_M(x) = \sum_{i=1}^{M} T(x, \Theta)$
- ***AdaBoost***: This paper uses AdaBoost, which originally applied to classification problem. And it incorporates the methodology of Bagging: choosing a subset of features every iteration.
- ***Parameters***:
  - Number of trees: $M_{GBT} = 100$.
  - Depth of trees: $J_{GBT} = 3$.
  - Learning rate: $\lambda_{GBT} = 0.1$.
  - Subset of features: $m_{GBT} = 15$.

Random Forests



- ***Bagging***: Using multiple sub-datasets to train many different models independently, and combining the predictions of these models to generate the ultimate results.
- ***Random Forests***: Consisting of many deep but decorrelated trees built on different samples of the data.
- ***Parameters***:
  - Number of trees: $B_{RAF} = 1000$.
  - Depth of trees: $J_{RAF} = 20$.
  - Subset of features: $m_{RAF} = \lfloor \sqrt{p} \rfloor$.

Ensembles

At period t, let $\hat{\mathcal{P}}^{s,a}_{t+1|t}$ denote the probability forecast of a learning algorithm a that stock s outperforms its cross-sectional median in period $t+1$, with $a \in \{DNN, GBT, RAF, ENS1, ENS2, ENS3\}$.

- ***Equal-weighted (ENS1)***: simple average.

$$\hat{\mathcal{P}}^{s,ENS1}_{t+1|t} = \frac{1}{3}(\hat{\mathcal{P}}^{s,DNN}_{t+1|t} + \hat{\mathcal{P}}^{s,GBT}_{t+1|t} + \hat{\mathcal{P}}^{s,RAF}_{t+1|t}) \tag{5}$$

- ***Performance-based (ENS2)***: Gini indices.

$$\begin{aligned} \hat{\mathcal{P}}^{s,ENS2}_{t+1|t} &= W^{DNN}_T \hat{\mathcal{P}}^{s,DNN}_{t+1|t} + W^{GBT}_T \hat{\mathcal{P}}^{s,GBT}_{t+1|t} + W^{RAF}_T \hat{\mathcal{P}}^{s,RAF}_{t+1|t} \\ W^i_T &= g^i_T/(g^{DNN}_T + g^{GBT}_T + g^{RAF}_T), i \in \{DNN, GBT, RAF\} \end{aligned} \tag{6}$$

- ***Rank-based (ENS3)***: Gini indices' ranks.

$$W^i_T = \frac{1/R^i_T}{1/R^{DNN}_T + 1/R^{GBT}_T + 1/R^{RAF}_T}, i \in \{DNN, GBT, RAF\} \tag{7}$$

Forecasting, ranking, and trading

- For each period $t+1$, forecast the probability $\hat{\mathcal{P}}^{s,a}_{t+1|t}$ for each stock s to outperform its cross-sectional median, with $a \in \{DNN, GBT, RAF, ENS1, ENS2, ENS3\}$ and $s \in \{1, \cdots, n_i\}$.
- Sorting all stocks in descending order separately by each forecast, results in six rankings – corresponding to the DNN, GBT, RAF, and ENS1, ENS2, ENS3.
- At the top, find the most undervalued stocks and at the bottom the most overvalued stocks in period $t+1$.
- In consequence, go long the top k stocks of each ranking, and short the bottom k stocks, with $k \in \{1, \cdots, n_i/2\}$. By censoring the middle part of the ranking, exclude the stocks with highest directional uncertainty from trading.
- Summary:
  - Trading frequency: daily.
  - Model training frequency: every study period (250 days).
  - Transaction costs: 0.05 percent per share per hald-turn.
  - Portfolio construction: Equal funds or equal amounts.
- We will also use other models, such as *XGBoost*, *LightGBM*, and *a more complicated NN*.
- We will use *cv* or *grid search* in the determining of the optimal parameters to make the result more robust.

General Results

This paper analyzes the performance of the portfolios consisting of the top k stocks, with $k \in \{1, \cdots, n_i/2\}$. They are compared in terms of *returns per day* prior to transaction costs, *standard deviation*, and *daily directional accuracy* at the portfolio level.

- The 3 ensembles all produce returns of approximately 0.45 percent per day, followed by RAF (0.43), GBT (0.37), GNN (0.33). Directional accuracy follows a similar pattern.
- Increasing k leads to decreasing returns and directional accuracy. (The latter indicator is always greater than 50%).
- In summary, the ensembles outperform all base models in terms of directional accuracy regardless of the level of k.
- In the following sections, we focus on the portfolio formed with $k = 10$. For the sake of simplicity, and the similarity in performance, we limit this analysis to the base learners and the equal-weighted ensemble (ENS1).

Strategy Performance Measure - Return and Risk Characteristics

Return Distribution	• **Mean Returns:** Average performance across strategies pre/post costs. • **Descriptive Statistics:** Minimum, Quartiles, Median, Maximum. • **Skewness and Kurtosis:** Analyzes asymmetry and fat tails in return distribution.
Risk Assessment	• **Standard Deviation:** Highlights overall strategy volatility. • **Value at Risk (VaR) and CVaR:** Quantifies potential maximum losses. • **Downside Deviation:** Focuses on downside risks. • **Maximum Drawdown:** Indicates potential loss in worst-case scenarios.
Risk-Return Characteristics	• **Sharpe Ratio:** Excess return per unit of total risk. • **Sortino Ratio:** Returns relative to downside risk. • **Calmar Ratio:** Returns relative to maximum drawdown.
Statistical Testing and Model Validation	• **t-Statistic(Newey-West):** Tests significance of mean returns, with adjustments for autocorrelation and heteroskedasticity. *H0: "Mean return equal to zero."* • **PT(Pesaran-Timmermann) Test Statistic :** Validates the predictive accuracy of the models. *H0: "Predictions and response are independently distributed."*

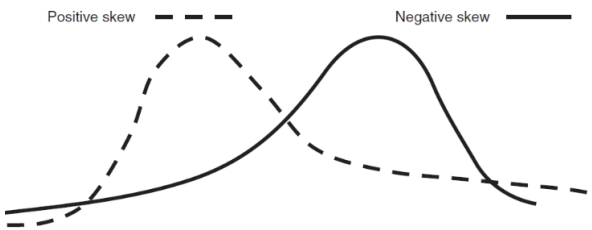
Strategy Performance Measure - Return and Risk Characteristics

Skewness: Measures the asymmetry of the return distribution.

- $$S_S = \sum \left(\frac{r_i - \bar{r}}{\sigma_{Sp}} \right)^3 \times \frac{n}{(n-1) \times (n-2)}$$
- *Financial Significance:* Positive skewness indicates a distribution with frequent small losses and occasional large gains.

Kurtosis: Measures the "tailedness" of the return distribution.

- $$K_S = \sum \left(\frac{r_i - \bar{r}}{\sigma_{Sp}} \right)^4 \times \frac{n \times (n+1)}{(n-1) \times (n-2) \times (n-3)}$$
- *Financial Significance:* High kurtosis (leptokurtic distribution) indicates a higher risk of extreme returns, both positive and negative, suggesting a higher potential for unexpected events ("fat tails"). Lower kurtosis (platykurtic distribution) suggests a distribution closer to normal distribution.



Kurtosis >3 peaked with fat tails

Strategy Performance Measure - Return and Risk Characteristics

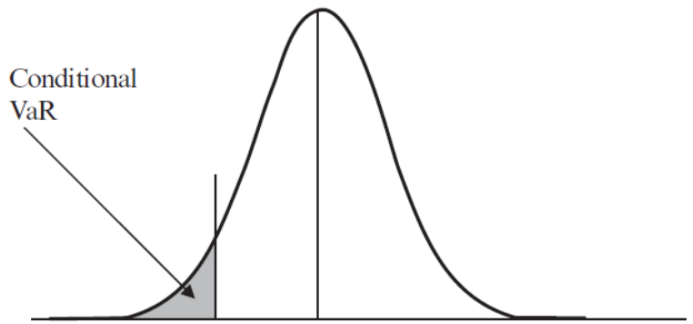
Value at Risk (VaR): Estimates maximum loss for a defined period at a given confidence level.

- $$P(r_P < -VaR_\alpha) = 1 - \alpha$$
$$VaR_\alpha = -(\bar{r} - z_\alpha \times \sigma)$$

where z_α is the z-score for confidence level α (based on a normal distribution of returns).

Conditional Value at Risk (CVaR), also known as Expected Shortfall: Measures the expected average loss beyond VaR.

- $$CVaR_\alpha = -\mathbb{E}[r_P | r_P < -VaR_\alpha]$$



Conditional VaR

- *Financial Significance:* Unlike VaR, which only provides the threshold level that losses are not expected to exceed with a certain confidence, CVaR estimates the average of all losses exceeding that threshold.

Strategy Performance Measure - Return and Risk Characteristics

Sharpe Ratio: Measures excess return per unit of risk.

- $$SR = \frac{r_P - r_F}{\sigma_P}$$
- *Financial Significance:* Indicates how much excess return is received for the extra volatility endured by holding a riskier asset. Higher values suggest better risk-adjusted performance.

Calmar Ratio: Measures the risk-adjusted return relative to the maximum drawdown.

- $$CR = \frac{r_P - r_F}{D_{Max}}$$
- *Financial Significance:* A high Calmar Ratio indicates a favorable risk-return profile, as it demonstrates higher returns relative to the risk of large losses.

Sortino Ratio: Similar to the Sharpe Ratio but only considers downside risk.

- $$SortinoRatio = \frac{r_P - r_F}{\sigma_D}$$
- *Financial Significance:* Provides insight into the risk of negative returns, focusing on downside deviation rather than total volatility, which is more relevant for investors concerned about downside risk.

Strategy Performance Measure - Model Validation (PT Test)

Indicator Variables

- Define indicator variables to translate economic predictions into binary outcomes

$$Y_t = 1 \text{ if } y_t > 0,\ 0 \text{ otherwise}; \quad X_t = 1 \text{ if } x_t > 0,\ 0 \text{ otherwise}; \quad Z_t = 1 \text{ if } y_t x_t > 0,\ 0 \text{ otherwise}.$$

Evaluation of predictive models:These variables help test the alignment of signs between predicted and actual economic changes, useful for evaluating predictive models.

Probability Estimates

- Estimate the probabilities of positive outcomes for y_t and x_t

$$P_y = \Pr(y_t > 0), \quad P_x = \Pr(x_t > 0).$$

- The proportion of correct predictions is given by $\hat{P} = n^{-1} \sum_{t=1}^{n} Z_t = \bar{Z}$, an empirical mean of correct predictions.

Hausman-Type Statistic

- Under the null hypothesis that y_t and x_t are independently distributed:

$$S_n = \frac{\hat{P} - \hat{P}_*}{\sqrt{\text{var}(\hat{P}) - \text{var}(\hat{P}_*)}} \sim N(0, 1)$$

- where $\hat{P}_* = \hat{P}_y \hat{P}_x + (1 - \hat{P}_y)(1 - \hat{P}_x)$, $\hat{P}_y = \bar{Y}$, $\hat{P}_x = \bar{X}$
- $\text{var}(\hat{P}) = n^{-1} \hat{P}_* (1 - \hat{P}_*)$
- $\text{var}(\hat{P}_*) = n^{-1} ((2\hat{P}_y - 1)^2 \hat{P}_x (1 - \hat{P}_x) + (2\hat{P}_x - 1)^2 \hat{P}_y (1 - \hat{P}_y)) + 4n^{-2} \hat{P}_y \hat{P}_x (1 - \hat{P}_y)(1 - \hat{P}_x)$

PT test results: A significant S_n suggests that x_t has predictive power over y_t, indicating dependency between the variables, contrary to the null hypothesis. This methodology extends beyond traditional sign tests by comparing the proportion of sign agreements rather than distributions, focusing on predictive accuracy.

Strategy Performance - Daily Return Characteristics

Return Performance

- Mean daily returns of 0.45% for equal-weighted ensemble, outperforming DNN (0.33%), GBT (0.37%), and RAF (0.43%) before transaction costs.
- Returns reduce to 0.25% per day after transaction costs.
- Newey–West t-statistics are significantly high (>12.0 before costs and >3.0 after costs) against a null hypothesis of zero mean return (critical value at 5% significance is 1.9600).

Model Predictional Accuracy

- PT test indicates significant predictional accuracy with statistics >13.0.

Risk and Return Characteristics

- Strategies show positive skewness and strong leptokurtosis, influenced by large outliers.
- Return contribution of long-leg ranges from 55% to 60% before costs, and 60% to 70% including costs.
- Historical 1-percent VaR is between -5.9% and -6.9%, roughly twice that of the general market, indicating substantial tail risk.

Tail Risk and Drawdowns

- RAFs show the lowest tail risk and DNNs the highest compared to other strategies.
- Maximum drawdowns are significant, with the equal-weighted ensemble at 74% and DNNs reaching 95%.
- Calmar ratio for the ensemble is 99%, showing better recovery potential compared to the market (17%).

Strategy Performance - Daily Return Characteristics

Daily return characteristics of $k = 10$ portfolio, prior to and after transaction costs for DNN, GBT, RAF, ENS1 compared to general market (MKT) from December 1992 until October 2015. NW denotes Newey–West standard errors with with one-lag correction and PT the Pesaran–Timmermann test.

	Before transaction costs				After transaction costs				
	DNN	GBT	RAF	ENS1	DNN	GBT	RAF	ENS1	MKT
Mean return (long)	0.0022	0.0025	0.0030	0.0029	0.0012	0.0015	0.0020	0.0019	–
Mean return (short)	−0.0011	−0.0013	−0.0013	−0.0015	−0.0001	−0.0003	−0.0003	−0.0005	–
Mean return	0.0033	0.0037	0.0043	0.0045	0.0013	0.0017	0.0023	0.0025	0.0004
Standard error (NW)	0.0004	0.0003	0.0003	0.0003	0.0004	0.0003	0.0003	0.0003	0.0001
t-statistic (NW)	8.6159	12.6786	14.9327	13.3962	3.3835	5.8952	7.9104	7.3781	2.8305
PT test statistic	14.7974	13.8185	15.5645	17.6052	–	–	–	–	–
Minimum	−0.1916	−0.1487	−0.1622	−0.1681	−0.1936	−0.1507	−0.1642	−0.1701	−0.0895
Quartile 1	−0.0079	−0.0062	−0.0053	−0.0063	−0.0099	−0.0082	−0.0073	−0.0083	−0.0046
Median	0.0025	0.0032	0.0033	0.0034	0.0005	0.0012	0.0013	0.0014	0.0008
Quartile 3	0.0133	0.0133	0.0127	0.0138	0.0113	0.0113	0.0107	0.0118	0.0058
Maximum	0.5475	0.2003	0.3754	0.4470	0.5455	0.1983	0.3734	0.4450	0.1135
Standard deviation	0.0269	0.0217	0.0208	0.0239	0.0269	0.0217	0.0208	0.0239	0.0117
Skewness	2.8547	0.2434	1.7283	2.6837	2.8547	0.2434	1.7283	2.6837	-0.1263
Kurtosis	50.5137	8.2267	29.8174	43.2698	50.5137	8.2267	29.8174	43.2698	7.9791
Historical 1-percent VaR	−0.0672	−0.0580	−0.0508	−0.0570	−0.0692	−0.0600	−0.0528	−0.0590	−0.0320
Historical 1-percent CVaR	−0.0929	−0.0831	−0.0721	−0.0786	−0.0949	−0.0851	−0.0741	−0.0806	−0.0461
Historical 5-percent VaR	−0.0322	−0.0262	−0.0229	−0.0267	−0.0342	−0.0282	−0.0249	−0.0287	−0.0179
Historical 5-percent CVaR	−0.0544	−0.0462	−0.0406	−0.0449	−0.0564	−0.0482	−0.0426	−0.0469	−0.0277
Maximum drawdown	0.5815	0.4391	0.3454	0.4017	0.9544	0.8425	0.6689	0.7367	0.5467
Calmar ratio	1.8884	3.2219	5.1033	4.6277	0.2813	0.5466	1.0037	0.9903	0.1692
Share with return > 0	0.5713	0.5939	0.6028	0.5883	0.5174	0.5351	0.5423	0.5367	0.5426

Daily Return Characteristics Conclusion

- Despite higher tail risks and drawdowns, the strategies offer substantial mean returns and quick recovery times, illustrating their potential value to investors seeking higher risk-adjusted returns.

Strategy Performance - Annualized Returns and Risk Measures

Annualized returns and risk measures of $k = 10$ portfolio, prior to and after transaction costs for DNN, GBT, RAF, ENS1 compared to general market (MKT) from December 1992 until October 2015.

	Before transaction costs				After transaction costs				
	DNN	GBT	RAF	ENS1	DNN	GBT	RAF	ENS1	MKT
Mean return	1.0981	1.4149	1.7627	1.8588	0.2685	0.4605	0.6714	0.7296	0.0925
Mean excess return	1.0446	1.3534	1.6924	1.7861	0.2361	0.4232	0.6287	0.6855	0.0646
Standard deviation	0.4277	0.3438	0.3308	0.3793	0.4277	0.3438	0.3308	0.3793	0.1852
Downside deviation	0.2474	0.2113	0.1847	0.2049	0.2615	0.2250	0.1981	0.2188	0.1307
Sharpe ratio	2.4426	3.9364	5.1166	4.7091	0.5521	1.2310	1.9008	1.8073	0.3486
Sortino ratio	4.4384	6.6956	9.5462	9.0702	1.0268	2.0466	3.3883	3.3337	0.7077

- **Annualized Returns (Post-Costs)**: ENS leads with 73%, followed by RAF (67%),GBT (46%), and DNN (27%).
- **Sharpe Ratio**: Equal-weighted ensemble's Sharpe ratio of 1.81 outperforms the general market significantly, indicating returns over ten times higher per unit of risk.
- **Comparative Sharpe Ratios**: Outperforms classical and generalized pairs trading, and competitive with advanced strategies:
  - Classical pairs trading: 0.59 (1962-2002)
  - Generalized pairs trading: 1.44 (1997-2007)
  - Deep conditional portfolio sorts: 2.96 (1968-2012, pre-costs)
  - Elman neural networks and ELECTRE III: 1.5 (1992-2006, pre-costs)
- **Sortino Ratio**: Reflects better risk management, with downside deviation for the ensemble at 0.22, about 1.7 times that of the general market.
- **Performance Ranking**: RAF performs best in risk-return metrics, followed by ENS, GBT, and DNN.

Strategy Performance Measure - Exposure to Systematic Sources of Risk

Factors represent different sources of systematic risk. They explain cross-sectional market anomalies by capturing the underlying risk dimensions and behavioral biases that drive deviations from the expected market efficiency.

Fama–French Three-Factor Model (FF3)

$$R_{i,t} - R_{f,t} = \alpha + \beta_{mkt}(R_{mkt,t} - R_{f,t}) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \epsilon_t$$

Factor Calculation

		B/M		
		High	Middle	Low
Size	Small	SH	SM	SL
	Big	BH	BM	BL

$$SMB = \frac{1}{3}(SH + SM + SL) - \frac{1}{3}(BH + BM + BL)$$

$$HML = \frac{1}{2}(SH + BH) - \frac{1}{2}(SL + BL)$$

Financial Significance

- SMB (Small Minus Big): SMB captures the historical premium of small-cap stocks over large-cap stocks.
- HML (High Minus Low): HML represents the performance advantage of value stocks (high book-to-market) over growth stocks (low book-to-market).

Strategy Performance Measure - Exposure to Systematic Sources of Risk

Factors represent different sources of systematic risk. They explain cross-sectional market anomalies by capturing the underlying risk dimensions and behavioral biases that drive deviations from the expected market efficiency.

Fama–French 3+2-Factor Model (FF3+2)

- Extends FF3 by including a momentum factor and a short-term reversal factor.
- $R_{i,t} - R_f = \alpha + \beta_{mkt}(R_{mkt,t} - R_f) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{Mom}Mom_t + \beta_{Rev}Rev_t + \epsilon_t$

Fama–French Five-Factor Model (FF5)

- Enhances the FF3 with two additional factors: robust minus weak profitability (RMW) and conservative minus aggressive investment (CMA).
- $R_{i,t} - R_f = \alpha + \beta_{mkt}(R_{mkt,t} - R_f) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{RMW}RMW_t + \beta_{CMA}CMA_t + \epsilon_t$

Fama–French VIX-Enhanced Model (FF VIX)

- Integrates the FF3+2 model with the VIX index (dummy variable), known as the "investor fear gauge".
- $R_{i,t} - R_f = \alpha + \beta_{mkt}(R_{mkt,t} - R_f) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{Mom}Mom_t + \beta_{Rev}Rev_t + \beta_{VIX}VIX_t + \epsilon_t$

Strategy Performance - Exposure to Systematic Sources of Risk

Equal-weighted ensemble strategy ENS1 with $k = 10$: exposure to systematic sources of risk after transaction costs from December 1992 until October 2015. Standard errors are depicted in parentheses.

	FF3	FF3+2	FF5	FF VIX
(Intercept)	0.0022***	0.0014***	0.0024***	0.0010**
	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Market	0.3271***	0.1759***	0.2172***	0.1903***
	(0.0269)	(0.0278)	(0.0300)	(0.0279)
SMB	−-0.0036	−0.0458		−0.0362
	(0.0524)	(0.0493)		(0.0492)
HML	−0.0290	0.2983***		0.3126***
	(0.0515)	(0.0515)		(0.0515)
Momentum		0.3885***		0.3972***
		(0.0355)		(0.0355)
Reversal		0.9474***		0.9387***
		(0.0361)		(0.0361)
SMB5			−0.0689	
			(0.0561)	
HML5			0.2348***	
			(0.0603)	
RMW5			−0.3308***	
			(0.0794)	
CMA5			−0.6639***	
			(0.0911)	
VIX				0.0047***
				(0.0010)
R^2	0.0259	0.1402	0.0381	0.1436
Adj. R^2	0.0254	0.1395	0.0373	0.1427
Num. obs.	5750	5750	5750	5750
RMSE	0.0236	0.0222	0.0234	0.0221

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

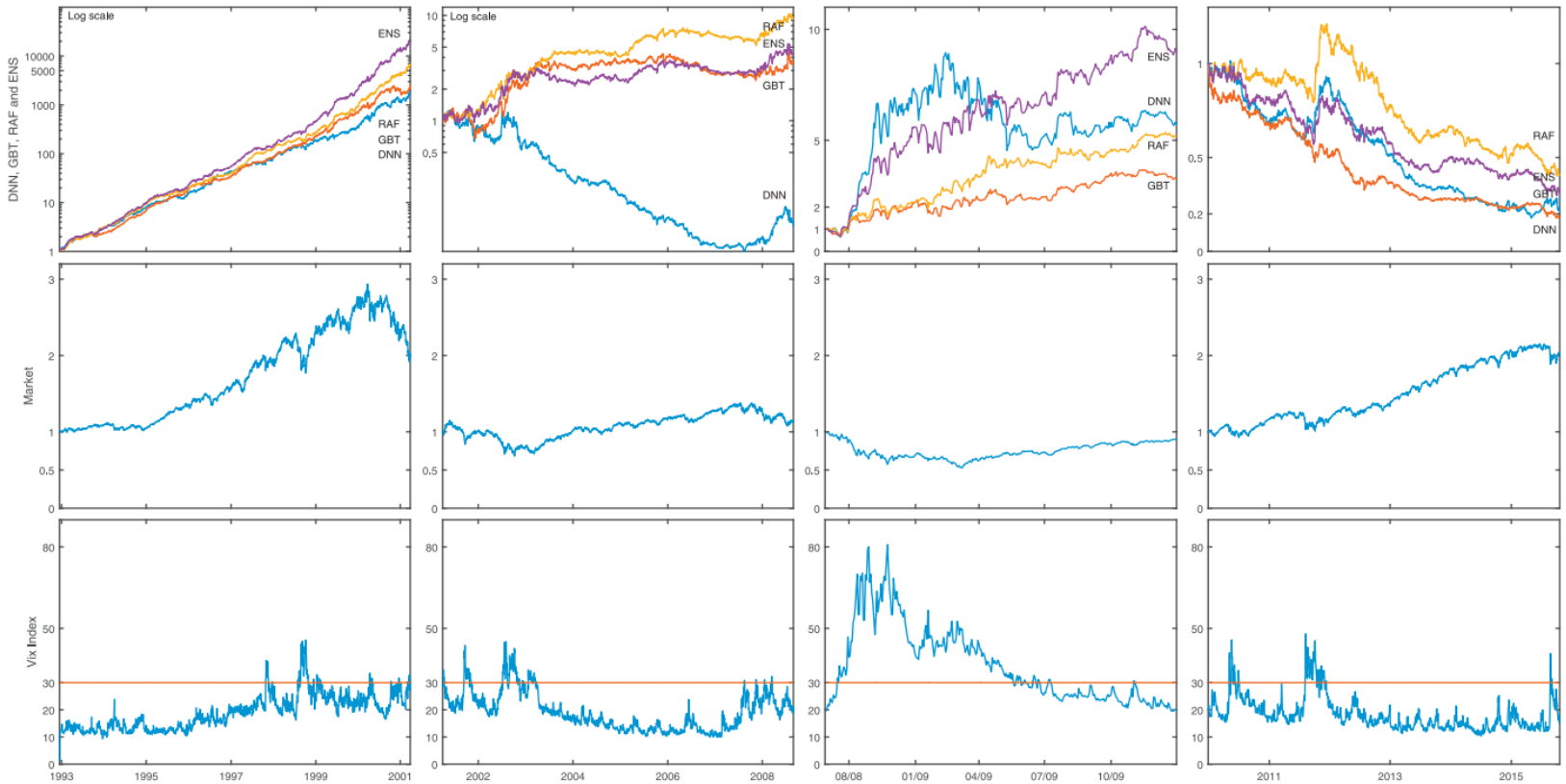
- **Daily Alpha:** Ranges from 0.14% to 0.24% across models.
- **Statistical Significance:** Strong in all models, especially high in FF3+2 and FF VIX. The ML algorithms extract momentum as well as shortterm reversal patterns from the data, thus explaining the factor loadings
- **Market Behavior:** Positive load on HML; negative on RMW and CMA indicating a tilt towards glamour stocks.
- **Volatility Sensitivity:** Enhanced performance during high market turmoil (VIX > 30).

Exposure to Systematic Sources of Risk Conclusion

- Strategy exhibits varied systematic risk loading, confirming significant alpha and sensitivity to market anomalies and volatility.

Sub-period Analysis

Sub-periods profile of k = 10 portfolio, after transaction costs for DNN, GBT, RAF, ENS1 compared to general market (MKT) and the VIX index from December 1992. until October 2015.



Strategy performance showing an initial phase of significant outperformance, followed by a moderation due to advances in machine learning, a spike in returns during the global financial crisis, and a subsequent decline as technological advances and increased algorithm accessibility reduced profitability.

Sub-period Analysis

Sub-Period 1 (12/92 - 03/01): Early Outperformance

- Period marked by strong and consistent outperformance before the widespread use of machine learning algorithms in finance.
- Techniques used today were not publicly available, unclear if used secretly in hedge funds.

Sub-Period 2 (04/01 - 08/08): Moderation Phase

- Introduction and popularization of powerful models like Random Forests and Stochastic Gradient Boosting.
- These models detect structures unseen before; however, post-introduction performance begins to deteriorate.

Sub-Period 3 (09/08 - 12/09): Global Financial Crisis

- All long-short strategies excel during this period of high market turmoil.
- Indicates robustness of strategies in volatile market conditions.

Sub-Period 4 (01/10 - 10/15): Deterioration of Returns

- Significant negative returns observed, with annualized rates between -14% and -25% after costs.
- Public accessibility and technological advancements lead to diminished profits; proprietary algorithms in hedge funds may still hold potential.

Further Analyses

Variable Importances

- ***Method:*** Utilizes H2O to analyze predictive strengths of features through weight matrices in deep learning and relative influence during tree splits in tree-based methods.
- ***Key Findings:*** Short-term returns (4-5 days) are the most predictive of market movements, while returns over 20 to 240 days show varied significance, with longer periods being more influential. Emphasizes the potential for trading strategies that capitalize on short-term anomalies for strategic gains.

Industry Breakdown

- ***Method:*** Analyzes S&P 500 industry representation from 1990 to 2015 using GICS codes compared to long and short positions in an equal-weighted ensemble strategy.
- ***Key Findings:*** Shows an overrepresentation in technology stocks and underrepresentation in financials, industrials, and utilities. Suggests a preference for high-beta over low-beta stocks in investment strategies.

Robustness Checks

- ***Method:*** Conducts sensitivity analysis on machine learning parameters to assess model stability and overfitting by varying the number of neurons and trees.
- ***Key Findings:*** DNN show consistent returns regardless of neuron count, highlighting benefits of feature abstraction and dropout regularization. Tree-based methods display slight return improvements with increased tree counts, indicating robustness without overfitting.

Conclusion: Methodology Overview

1. **Data Preparation**
   - *Historical Data Compilation:* Aggregate daily return data for all constituents of the S&P 500 index.
   - *Survivor Bias Correction:* Include past constituent lists to account for survivor bias.
2. **Feature Generation**
   - *Lagged Returns Calculation:* Determine lagged returns for periods up to 240 days to form input features.
   - *Binary Response Variable:* Create a binary outcome that indicates if a stock's return is above the market median.
3. **Model Training**
   - *Sequential Training and Trading Sets:* Use a rolling window for training sets of 750 days followed by trading sets of 250 days.
   - *Non-Overlapping Trading Periods:* Ensure each 250-day trading period is non-overlapping for valid out-of-sample testing.

Conclusion: Methodology Overview

4. **Ensemble Strategy:**
   - *ENS1 (Equal-Weighted Ensemble):* Equally average predictions from DNN, GBT, and RAF models.
   - *ENS2 (Performance-Based Ensemble):* Weight predictions based on Gini indices from the training period.
   - *ENS3 (Rank-Based Ensemble):* Assign weights based on model performance ranks during training.
5. **Out-of-Sample Forecasting and Trading:**
   - *Daily Probability Forecasting:* Forecast daily probability of each stock outperforming the market.
   - *Daily Stock Ranking:* Rank stocks based on predicted probabilities.
   - *Positioning:* Go long on the top k stocks and short on the bottom k, excluding middle-ranked stocks.
6. **Trading Frequency and Horizon:**
   - *Daily Rebalancing:* Rebalance portfolio daily according to the latest model forecasts.
   - *One-Day Ahead Prediction Horizon:* Generate trading signals for a one-day ahead period.

Conclusion: Possible Improvement

- *Using other models, such as XGBoost, LightGBM, and a more complicated NN.*
- *Using cv or grid search in the determining of the optimal parameters to make the results more robust.*
- *Incorporating nonlinearities and interaction terms of features into the predictive models, to capture complex relationships within the feature space.*
- *Generating more features based on price and volume data, to capture more sophisticated relationships existing in the market.*
- *Implementing CVaR-portfolio optimization to mitigate risks.*
- *Investigating advanced ensemble integration methods such as adaptive weighting to potentially enhance performance.*
- *Expanding the time range to 1990-2023.*