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

Final result

RunPing Zhao 211232011
Haohan Zhao 211098172

Nanjing University

June 5, 2024

1. Introduction
2. Data and Software
3. Methodology
4. Result
5. Possible Improvement

1. Introduction
2. Data and Software
3. Methodology
4. Result
5. Possible Improvement

Introduction

- 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.
- Generally speaking, the arbitrage strategy is to predict the probability of a stock to outperform the general market, thus generating the daily one-day-ahead trading signals. Then long the highest k and short the lowest k stocks.
- Our work is to replicate the methodology of this paper on the newest market data, and replace GBDT with some more powerful boosting trees like ***XGBoost*** and ***LightGBM***.
- We conduct data processing, model training and predicting, and holistic performance testing. The results are presented below.

1. Introduction
2. Data and Software
3. Methodology
4. Result
5. Possible Improvement

Data and Software

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

1. The lists of S&P 500 components since 1996 (update for 2024-04-08), along with their respective industries and historical changes, are retrieved from this ***GitHub repository***. The repository is copyrighted by Farrell J. Aultman (fja0568@gmail.com) and is maintained regularly.
2. Eliminate survivor bias following Krauss and Stübinger (2015).
3. Obtain all month end constituent lists for the S&P 500 from December 1995 to March 2024, then consolidate these lists into one binary matrix, indicating whether the stock is a constituent of the index in the subsequent month or not.
4. For all stocks having ever been a constituent of the index, download the daily adjusted close data from January 1996 until April 2024 from ***yfinance*** in Python.
5. The total number of stocks is 1163, excluding the stocks that have been delisted and cannot obtain data, and the number of effective trading days is less than 1000, the remaining 791 stocks.

   Since there is no publicly available official constituents history data, we have searched many databases, but the maintenance of each database is different, and the delisting/renaming/acquisition of the company is not effectively handled, so the data quality is relatively poor. Our data is somewhat different from the paper's original data, but we have tried our best to clean the data and make it reasonable.

All of our work are based on ***Python***. DNN is constructed using ***Pytorch***, GBDT and Random Forests use ***sklearn***, XGBoost and LightGBM use ***xgboost*** and ***lightgbm*** in Python.

1 Introduction

2 Data and Software

3 Methodology

Generation of Training and Trading Sets

Feature Engineering

Deep Neural Network

Random Forests

Gradient Boosting Decision Tree

eXtreme Gradient Boosting

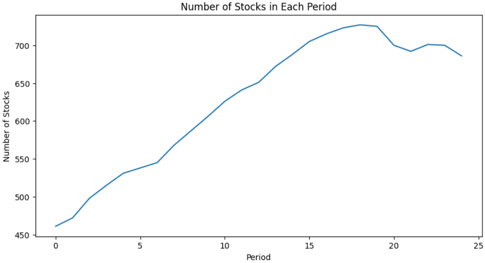
Light Gradient Boosting Machine

Ensemble

4 Result

Split Trainig 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).
- The split of study period uses rolling window with 250 as window size, which makes all trading sets continuous but non-overlapping.
- Our data has 7130 tradings days so we drop the first 130 days and split the remaining data into 25 study periods. So actually our training data began at July 5, 1996.
- In every study period, we use 250 days as trading set, and 240 days to compute features, so we drop the stocks whose effective trading days are less than 600. The final stock numbers of all periods are shown below.



Generating Features

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

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

$$R^s_{t,m} = \log\left(\frac{P^s_t}{P^s_{t-m}}\right), \quad m \in \{\{1, \cdots, 20\} \cap \{40, 60, \cdots, 240\}\} \quad (1)$$

- ***Target:*** A binary variable $Y^s_{t+1} \in \{0, 1\}$ for stock s, which equals 1 when its return $R^s_{t+1,1}$ is greater than market median return.
- Purpose: Forecast a probability $\mathscr{P}^s_{t+1|t}$ for stock s to outperform the cross-sectional median in period $t + 1$.
- Training days: 510; Trading days: 250.
- Training sets: $510 \times n_i \times 32$; Trading sets: $250 \times n_i \times 31$[1].

[1] n_i is the stock number of the i_{th} study period, where $i \in \{1, 2, \cdots, 25\}$

Deep Neural Network

- ***Topology of our net:*** I–H1–H2–H3–O with 31-31-10-5-2 neurons, 2746 parameters in total, so 93 training examples per parameter (when $n_i = 500$).
- All layers are composed of many neurons, which will output a weighted combination of n_l outputs of layer l, plus a bias b:

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

- ***Activation Function:*** MAXOUT for hidden layers:

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

- ***Activation Function:*** Sigmoid for output layer:

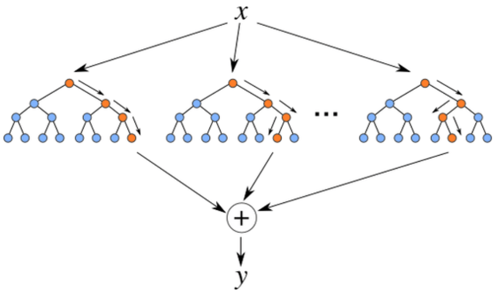
$$sigmod(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

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

$$\mathscr{L}(W, B|j) = -\sum_{y \in \mathcal{O}} \left(ln(o_y^{(j)}) t_y^{(j)} + ln(1 - o_y^{(j)})(1 - t_y^{(j)}) \right) \tag{5}$$

- ***Optimization:*** Adam optimizer in Pytorch with learning rate 0.001.
- ***Regularization:*** *Dropout* with drop out rate 0.1 for input layer and 0.5 for hidden layers; *L1* with shrinkage rate $\lambda_{DNN} = 10^{-5}$.
- ***Epochs:*** 400, with an early stopping strategy (if the training loss dose not decrease for the last 20 epochs).

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: $m_{RAF} = \lfloor \sqrt{p} \rfloor$.

Gradient Boosting Decision Tree

- ***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: $f_M(x) = \sum_{m=1}^{M} \rho T_m(x, \Theta)$
- ***GBDT:*** We use GBDT, which is proficient in classification problem while using classification tree as base learner. And we 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: $m_{GBT} = 15$.

XGBoost

- In machine learning, bias and variance is two common problems, which measure the accuracy and stability, respectively.
- XGBoost is an improved version of GBDT, improves model's accuracy by ***adding second derivative of loss function*** and reduce model's variance by ***imposing penalty on model's parameters***. The objective function of XGBoost can be expressed below:

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant
\end{aligned}
\tag{6}
$$

$$
where \quad l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) = l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)
$$

- The objective function is the most important difference between XGBoost and GBDT, and it's this difference makes it perform better on both bias and variance.
- XGBoost also use many other methods such as *column block parallel learning*, special *optimal segmentation algorithm* and so forth to improve training efficiency.
- ***parameters:***
  - learning rate: 0.1
  - max depth: 5
  - l2 regularization rate: 0.1
  - subsample and subfeatures: 0.75

LightGBM

LightGBM is an improved version of XGBoost, which performs better on large scale datasets for its training efficiency but still high accuracy.
The overall training complexity of XGBoost can be roughly estimated as:

number of trees × number of leaves of each tree × the complexity of generate every leaf

Since the base model adopted by XGBoost is a binary tree, each leaf needs to be split once to generate. For each split, all candidate split points on all features need to be traversed to calculate the maximum improvement. So the complexity of generating a leaf node can be estimated as:

number of features × number of candidate split points × number of samples

LightGBM mainly uses three methods to reduce the complexity of generating a leaf node: ***Histogram*** to reduce the number of candidate split points, ***GOSS*** to reduce the number of samples, and ***EFB*** to reduce the number of features.

- ***Histogram:*** It works by discretizing continuous features into a fixed number of bins, such as 255, to reduce the number of candidate split points to $N_{bin} - 1$.
- ***Gradient-based One-Side Sampling:*** It retains the samples with larger absolute gradient values while sample those with smaller gradient values in a certain proportion.
- ***Exclusive Feature Bundling:*** It combines the sparse features (most of their values are 0 so they are thought to be mutually exclusive) together.

Parameters: learning rate: 0.05; num_leaves: 31; subsample and subfeatures: 0.8.

Ensemble

Now we have constructed DNN, RAF, GBDF, XGBoost and LightGBM models. Then we consider the ensembles of these models.

We devide these five models into two groups $G_1 = \{DNN, RAF\}$ and $G_2 = \{GBDT, XGB, LGB\}$ first based on their intrinsic characteristics, where DNN is a neural network, RAF uses the method of bagging, and the other three use the method of boosting.

We fix G_1, and choose one form G_2 to compose an ensemble model, and use three different methods to weight three models' predictions, thus we get 9 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 $a \in \{DNN, RAF, GBT, XGB, LGB, ENS1 \sim 9\}$.

- ***Equal-weighted***: simple average.

$$\hat{\mathcal{P}}^{s,equal}_{t+1|t} = \frac{1}{3}(\hat{\mathcal{P}}^{s,a_i}_{t+1|t} + \hat{\mathcal{P}}^{s,a_j}_{t+1|t} + \hat{\mathcal{P}}^{s,a_k}_{t+1|t}) \tag{7}$$

- ***Performance-based***: Gini indices.

$$\begin{gathered} \hat{\mathcal{P}}^{s,gini}_{t+1|t} = W^{a_i}_T \hat{\mathcal{P}}^{s,a_i}_{t+1|t} + W^{a_j}_T \hat{\mathcal{P}}^{s,a_j}_{t+1|t} + W^{a_k}_T \hat{\mathcal{P}}^{s,a_k}_{t+1|t} \\ W^m_T = g^m_T/(g^{a_i}_T + g^{a_j}_T + g^{a_k}_T), m \in \{a_i, a_j, a_k\} \end{gathered} \tag{8}$$

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

$$W^m_T = \frac{1/R^m_T}{1/R^{a_i}_T + 1/R^{a_j}_T + 1/R^{a_k}_T}, m \in \{a_i, a_j, a_k\} \tag{9}$$

1 Introduction

2 Data and Software

3 Methodology

4 Result
- General Results
- Strategy performance
- Sub-period analysis
- Further analyses

5 Possible Improvement

General Results

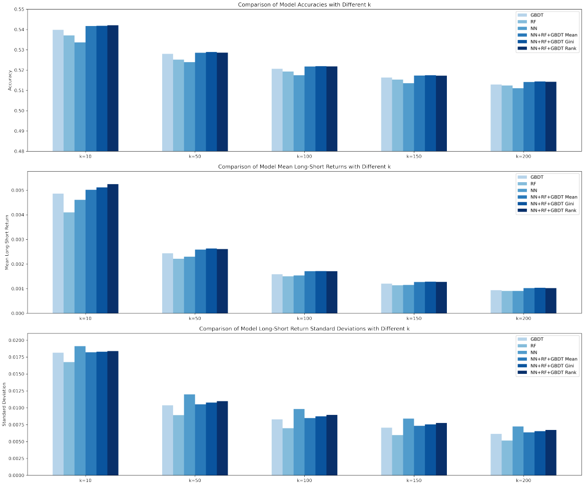


Fig. 1. Daily performance metrics for long-short portfolios of different sizes: mean return, standard deviation, and directional accuracy from July 1999 until December 2023.

General Results

Analyzed portfolios of top k stocks (k ∈ 10, 50, 100, 150, 200). Compared in terms of daily ***returns before transaction costs, standard deviation, and daily directional accuracy***.

Main Findings:

- Directional accuracy and returns decline as k increases.
  - Directional accuracy almost aligns with the pattern of returns.
  - Increasing k results in decreasing standard deviations. This aligns with classical portfolio theory.
- Base Learner Comparison: GBT> DNN > RAF(original RAF>GBT>DNN).
- ENS1(Equal-Weighted), ENS2(Performance-Based), and ENS3(Rank-Based) show similar performance(0.50% returns per day).Differences are less than 0.01% per day.

Further analysis focus on **k = 10** portfolio and limited to **base learners and ENS1**(equal-weighted ensemble). Results presented **before and after transaction costs of 0.05% per share per half-turn**.

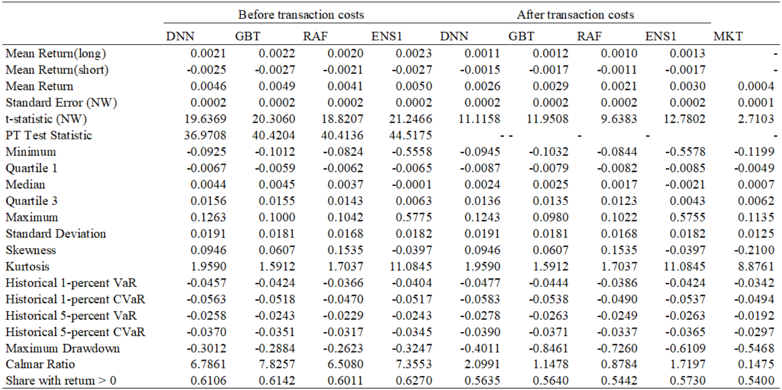
Explanation on Backtesting Metric Calculations

Explanation on Backtesting Metric Calculations

- The evaluation metrics for risk, return, and risk-adjusted return in the backtest use the `empyrical`, `scipy.stats`, and `statsmodels.api` libraries in Python.
- VaR and CVaR were calculated using historical quantiles.
- Pesaran-Timmermann-Test (a hypothesis test including the Hausman statistic) function was written by ourselves (*ref. link: GitHub Link*).
- Newey-West standard error and t-statistic calculation were written by ourselves.(*ref. link: CSDN Article*).
- We thank French for providing all relevant data for asset pricing models (FF3, FF5, momentum and short term reversal factors) on his website (*ref. link:Ken French Data Library*).
- VIX data downloaded from the website of Historical Data for Cboe VIX® Index (*ref. link:Cboe VIX Historical Data*).

Daily Return Characteristics

Table 2 Daily return characteristics of k = 10 portfolio, prior to and after transaction costs for DNN, GBT, RAF, ENS1 compared to general market (MKT) from July 1999 until December 2023. 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)	0.0021	0.0022	0.0020	0.0023	0.0011	0.0012	0.0010	0.0013	-
Mean Return(short)	-0.0025	-0.0027	-0.0021	-0.0027	-0.0015	-0.0017	-0.0011	-0.0017	-
Mean Return	0.0046	0.0049	0.0041	0.0050	0.0026	0.0029	0.0021	0.0030	0.0004
Standard Error (NW)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001
t-statistic (NW)	19.6369	20.3060	18.8207	21.2466	11.1158	11.9508	9.6383	12.7802	2.7103
PT Test Statistic	36.9708	40.4204	40.4136	44.5175	- -	-	-		-
Minimum	-0.0925	-0.1012	-0.0824	-0.5558	-0.0945	-0.1032	-0.0844	-0.5578	-0.1199
Quartile 1	-0.0067	-0.0059	-0.0062	-0.0065	-0.0087	-0.0079	-0.0082	-0.0085	-0.0049
Median	0.0044	0.0045	0.0037	-0.0001	0.0024	0.0025	0.0017	-0.0021	0.0007
Quartile 3	0.0156	0.0155	0.0143	0.0063	0.0136	0.0135	0.0123	0.0043	0.0062
Maximum	0.1263	0.1000	0.1042	0.5775	0.1243	0.0980	0.1022	0.5755	0.1135
Standard Deviation	0.0191	0.0181	0.0168	0.0182	0.0191	0.0181	0.0168	0.0182	0.0125
Skewness	0.0946	0.0607	0.1535	-0.0397	0.0946	0.0607	0.1535	-0.0397	-0.2100
Kurtosis	1.9590	1.5912	1.7037	11.0845	1.9590	1.5912	1.7037	11.0845	8.8761
Historical 1-percent VaR	-0.0457	-0.0424	-0.0366	-0.0404	-0.0477	-0.0444	-0.0386	-0.0424	-0.0342
Historical 1-percent CVaR	-0.0563	-0.0518	-0.0470	-0.0517	-0.0583	-0.0538	-0.0490	-0.0537	-0.0494
Historical 5-percent VaR	-0.0258	-0.0243	-0.0229	-0.0243	-0.0278	-0.0263	-0.0249	-0.0263	-0.0192
Historical 5-percent CVaR	-0.0370	-0.0351	-0.0317	-0.0345	-0.0390	-0.0371	-0.0337	-0.0365	-0.0297
Maximum Drawdown	-0.3012	-0.2884	-0.2623	-0.3247	-0.4011	-0.8461	-0.7260	-0.6109	-0.5468
Calmar Ratio	6.7861	7.8257	6.5080	7.3553	2.0991	1.1478	0.8784	1.7197	0.1475
Share with return > 0	0.6106	0.6142	0.6011	0.6270	0.5635	0.5640	0.5442	0.5730	0.5400

Daily Return Characteristics

Return Performance

- ENS1(0.50%, 0.03% after transaction costs)>GBT(0.49%)>DNN(0.46%)>RAF(0.41%). *original paper: Mean daily returns of 0.45% for equal-weighted ensemble, outperforming DNN (0.33%), GBT (0.37%), and RAF (0.43%) before transaction costs.*
- Short position(51%-54%) contributes more than the long position(41%-47%).
- Skewness and kurtosis disappear due to winsorization (5%) and the use of logarithmic returns.

Risk Characteristics

- Historical 1-percent VaR is between -4.0% and -4.2%, bigger than that of the general market(-3.4%), indicating tail risk. *original paper:Historical 1-percent VaR is between -5.9% and -6.9%.*
- Maximum drawdowns are significant, with GBT at 84% and RAF reaching 73%.
- Calmar ratio for the ensemble is 171%, showing better recovery potential compared to the market (14%). *original paper:Calmar ratio for the ensemble is 99%*

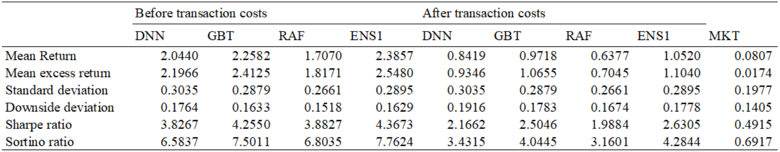
Model Accuracy:

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

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.

Annualized Returns and Risk Measures

Table 3 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 July 1999 until December 2023.



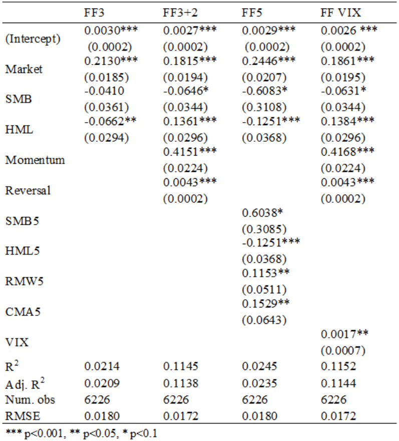
	Before transaction costs				After transaction costs				
	DNN	GBT	RAF	ENS1	DNN	GBT	RAF	ENS1	MKT
Mean Return	2.0440	2.2582	1.7070	2.3857	0.8419	0.9718	0.6377	1.0520	0.0807
Mean excess retum	2.1966	2.4125	1.8171	2.5480	0.9346	1.0655	0.7045	1.1040	0.0174
Standard deviation	0.3035	0.2879	0.2661	0.2895	0.3035	0.2879	0.2661	0.2895	0.1977
Downside deviation	0.1764	0.1633	0.1518	0.1629	0.1916	0.1783	0.1674	0.1778	0.1405
Sharpe ratio	3.8267	4.2550	3.8827	4.3673	2.1662	2.5046	1.9884	2.6305	0.4915
Sortino ratio	6.5837	7.5011	6.8035	7.7624	3.4315	4.0445	3.1601	4.2844	0.6917

Annualized Returns and Risk Measures

- ***Annualized Returns (Post-Costs)***: ENS leads with 105%, followed by GBT (97.2%), DNN(84.2%), and RAF(63.8%).
  Original papar: ENS leads with 73%, followed by RAF (67%),GBT (46%), and DNN (27%).
- ***Sharpe Ratio***:
  - Equal-weighted ensemble's Sharpe ratio of 4.28 outperforms the general market significantly, indicating returns over 5.36(2.63/0.49) times higher per unit of risk.
  - Comparative Research: 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 0.16, about 6.20(4.28/0.69) times that of the general market.
- ***Risk-eturn Performance Ranking***: ENS>GBT>RAF>DNN.

Exposure to Systematic Sources of Risk

Table 4 Equal-weighted ensemble strategy ENS1 with k = 10 : exposure to systematic sources of risk after transaction costs from July 1999 until December 2023. Standard errors are depicted in parentheses.



	FF3	FF3+2	FF5	FF VIX
(Intercept)	0.0030*** (0.0002)	0.0027*** (0.0002)	0.0029*** (0.0002)	0.0026 *** (0.0002)
Market	0.2130*** (0.0185)	0.1815*** (0.0194)	0.2446*** (0.0207)	0.1861*** (0.0195)
SMB	-0.0410 (0.0361)	-0.0646* (0.0344)	-0.6083* (0.3108)	-0.0631* (0.0344)
HML	-0.0662** (0.0294)	0.1361*** (0.0296)	-0.1251*** (0.0368)	0.1384*** (0.0296)
Momentum		0.4151*** (0.0224)		0.4168*** (0.0224)
Reversal		0.0043*** (0.0002)		0.0043*** (0.0002)
SMB5			0.6038* (0.3085)	
HML5			-0.1251*** (0.0368)	
RMW5			0.1153** (0.0511)	
CMA5			0.1529** (0.0643)	
VIX				0.0017** (0.0007)
R^2	0.0214	0.1145	0.0245	0.1152
Adj. R^2	0.0209	0.1138	0.0235	0.1144
Num. obs	6226	6226	6226	6226
RMSE	0.0180	0.0172	0.0180	0.0172

*** p<0.001, ** p<0.05, * p<0.1

Exposure to Systematic Sources of Risk

- ***Daily Alpha:*** Ranges from 0.26% to 0.30%.Statistical Significant in all models. *orginal paper: from 0.14% to 0.24%* Ranges from 0.14% to 0.24% across models.
- ***Momentum and short-term reversal patterns:***The ML algorithms extract momentum as well as shortterm reversal patterns from the data, *which is consistent with the original paper.*
- ***Market Behavior:*** The positivity or negativity of the loadings on the five factors are inconsistent with original paper, and the models are not consistent with each other, and thus market behavior is not very certain..
- ***Volatility Sensitivity:*** Enhanced performance during high market turmoil (VIX > 30),*which is consistent with the original paper*.

Strategy exhibits varied systematic risk loading, confirming significant alpha and sensitivity to market anomalies and volatility, which is consistent with the original paper.

Sub-periods Analysis

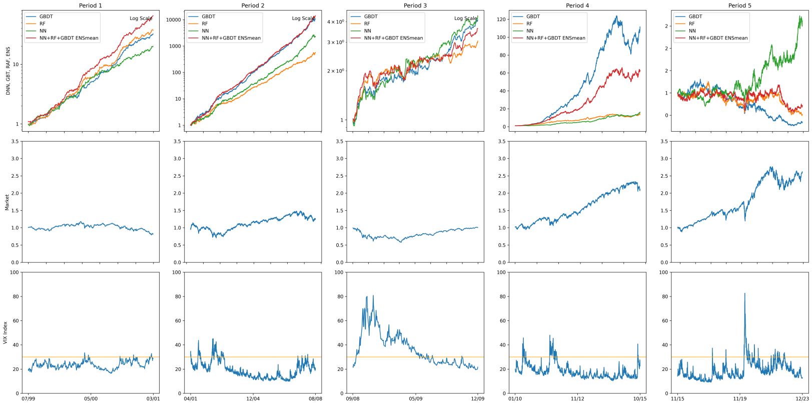


Fig. 2. Sub-periods profile of k = 10 portfolio, after transaction costs for DNN, GBT, RAF, ENS1 compared to general market (MKT) and the VIX index from July 1999 until December 2023.

Sub-periods Analysis

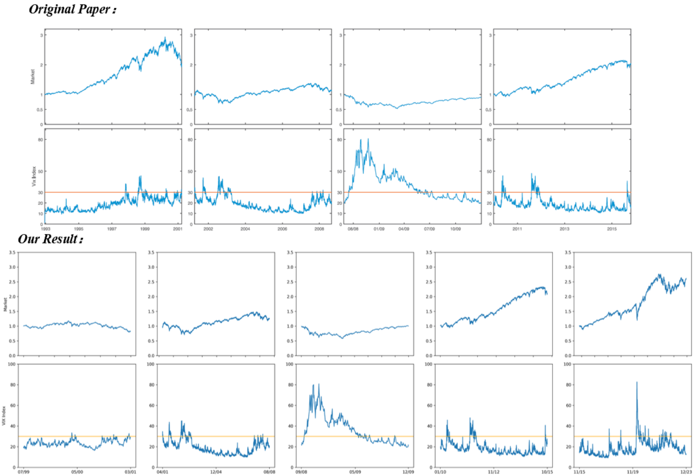
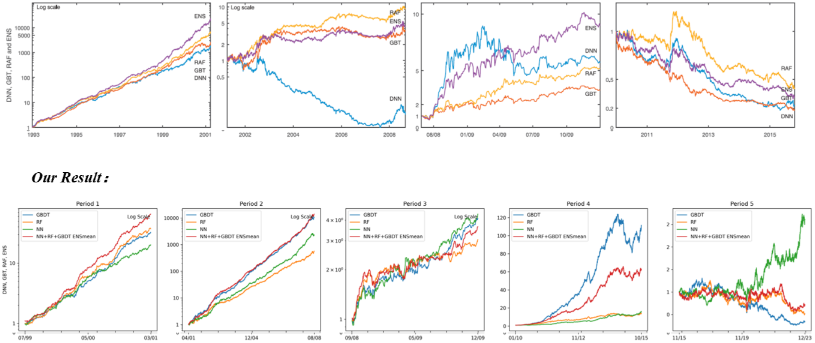


Fig. 2.1 Compairson of original paper and our result (general market (MKT) and the VIX index from July 1999 until December 2023.)

Sub-periods Analysis

Original Paper :



Our Result :

Fig. 2.2 Compairson of original paper and our result (sub-periods profile of k = 10 portfolio, after transaction costs for DNN, GBT, RAF, ENS1 from July 1999 until December 2023.)

Sub-period Analysis

Sub-Period 1 (07/99 - 04/08): 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.

Our model shows no "Moderation Phase", which is "04/01 - 08/08" in original paper.

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

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

Sub-Period 4 (01/10 - 10/15): Moderation Phase

- It proves the introduction and popularization of powerful models which detects structures unseen before like Stochastic Gradient Boosting.
- Post-introduction performance like NN and Random Forests begins to deteriorate, outperformed by GBT.

Sub-Period 5 (11/15 - 12/23): Deterioration of Returns

- Significant negative returns observed, with annualized rates between -7% and -15% after costs.
- Public accessibility and technological advancements lead to diminished profits; proprietary algorithms in hedge funds may still hold potential.

Sub-period Analysis

Table 5 Annualized risk-return characteristics per sub-period for DNN, GBT, RAF, ENS1.

		Before transaction costs				After transaction costs				
		DNN	GBT	RAF	ENS1	DNN	GBT	RAF	ENS1	MKT
Period 07/99-03/01										
	Mean return	7.5049	10.2080	11.1282	14.4298	4.1566	5.7993	6.3588	9.0945	-0.0808
	Mean excess return	7.9396	10.7812	11.6946	15.1159	4.4214	6.1486	6.7040	9.9848	0.0550
	Standard deviation	0.3621	0.3489	0.3331	0.3596	0.3621	0.3489	0.3331	0.3652	0.2278
	Sharpe ratio	6.1159	7.1328	7.6932	7.8280	4.7241	5.6882	6.1802	6.5402	-0.2563
	Maximum drawdown	-0.0821	-0.1179	-0.0844	-0.0815	-0.0932	-0.1338	-0.0900	-0.0872	-0.3080
	Calmar ratio	91.4134	86.6181	131.9135	177.0372	44.5915	43.3339	70.6483	104.3364	-0.2625
Period 04/01-08/08										
	Mean return	3.4905	4.5655	2.7763	4.7873	1.7193	2.3718	1.2861	2.5025	0.0435
	Mean excess return	3.6849	4.7739	2.9015	5.0080	1.8375	2.4985	1.3622	2.6276	0.0271
	Standard deviation	0.2932	0.2709	0.2474	0.2676	0.2932	0.2709	0.2474	0.2676	0.1697
	Sharpe ratio	5.2831	6.4935	5.5083	6.7160	3.5643	4.6328	3.4710	4.8287	0.3357
	Maximum drawdown	-0.1971	-0.1684	-0.1567	-0.1296	-0.2181	-0.1835	-0.1836	-0.1493	-0.3870
	Calmar ratio	17.7068	27.1184	17.7175	36.9418	7.8827	12.9240	7.0059	16.7558	0.1124
Period 09/08-12/09										
	Mean return	3.5968	3.3903	2.5852	3.1220	1.7836	1.6583	1.1700	1.4279	-0.0705
	Mean excess return	3.9539	3.7419	2.7966	3.4620	2.0008	1.8721	1.2985	1.6268	0.0027
	Standard deviation	0.3903	0.3953	0.3409	0.4011	0.3903	0.3953	0.3409	0.3991	0.3932
	Sharpe ratio	4.1125	3.9494	3.9239	3.7393	2.8212	2.6744	2.4454	2.4245	0.0103
	Maximum drawdown	-0.1629	-0.1215	-0.1503	-0.2087	-0.1798	-0.1460	-0.1905	-0.2310	-0.4637
	Calmar ratio	22.0747	27.9146	17.1954	14.9558	9.9184	11.3547	6.1426	6.1817	-0.1521
Period 01/10-10/15										
	Mean return	1.6020	2.6293	1.5450	2.2725	0.5740	1.1969	0.5395	0.9893	0.1241
	Mean excess return	1.6917	2.7725	1.6425	2.4021	0.6285	1.2839	0.5987	0.9242	0.0004
	Standard deviation	0.2532	0.2568	0.2474	0.2562	0.2532	0.2568	0.2474	0.2568	0.1643
	Sharpe ratio	3.9100	5.1594	3.9052	4.7656	1.9195	3.1970	1.8684	2.8095	0.7946
	Maximum drawdown	-0.1392	-0.1915	-0.1715	-0.1440	-0.1923	-0.3582	-0.2653	-0.2476	-0.2033
	Calmar ratio	11.5074	13.7297	9.0073	15.7812	2.9854	3.3411	2.0337	3.9953	0.6105
Period 11/15-12/23										
	Mean return	0.7749	0.3966	0.4760	0.5347	0.0729	-0.1562	-0.1081	-0.0772	0.1212
	Mean excess return	0.8895	0.4692	0.5410	0.6086	0.1423	-0.1122	-0.0687	-0.0805	0.0140
	Standard deviation	0.3109	0.2778	0.2578	0.2812	0.3109	0.2778	0.2578	0.2812	0.1909
	Sharpe ratio	2.0024	1.3421	1.6403	1.6648	0.3815	-0.4719	-0.3150	-0.1449	0.6953
	Maximum drawdown	-0.3012	-0.2884	-0.2623	-0.3247	-0.4011	-0.8461	-0.6890	-0.6059	-0.3425
	Calmar ratio	2.5728	1.3752	1.8147	1.6470	0.1816	-0.1846	-0.1570	-0.1274	0.3538

Sub-period Analysis

To summarize,

- *The performance of our model is generally lagging behind the original paper, especially inconsistent with 10-15 year period. It shows signs of weakness after 2015, but to some extent, it is consistent with the original paper.*
- *Strategy performance showing an initial phase of significant outperformance, a spike in returns during the global financial crisis, followed by a moderation due to advances in machine learning, and a subsequent decline as technological advances and increased algorithm accessibility reduced profitability.*

Variable Importances

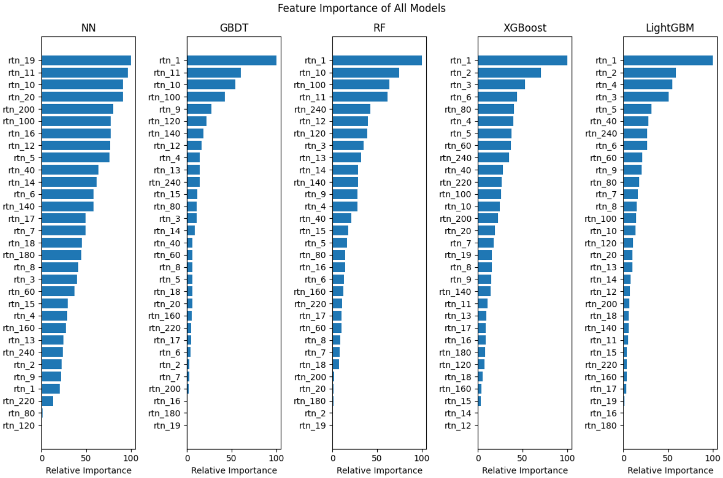
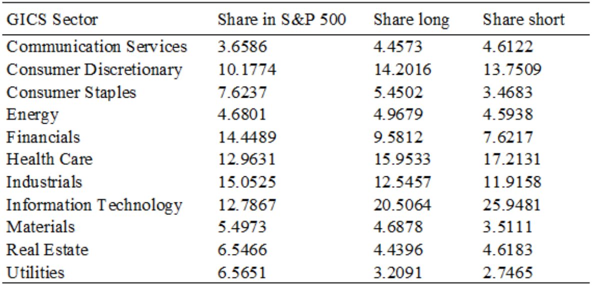


Fig. 3. Variable importance extracted from DNN, GBT, RAF July 1999 until December 2023. Most important variable normalized to 100.

Industry Breakdown

Table 6 Equal-weighted ensemble strategy with k = 10 from July 1999 until December 2023: breakdown of S&P 500 constituents by industry versus breakdown of ENS1 long and short portfolio holdings by industry, in percent.



GICS Sector	Share in S&P 500	Share long	Share short
Communication Services	3.6586	4.4573	4.6122
Consumer Discretionary	10.1774	14.2016	13.7509
Consumer Staples	7.6237	5.4502	3.4683
Energy	4.6801	4.9679	4.5938
Financials	14.4489	9.5812	7.6217
Health Care	12.9631	15.9533	17.2131
Industrials	15.0525	12.5457	11.9158
Information Technology	12.7867	20.5064	25.9481
Materials	5.4973	4.6878	3.5111
Real Estate	6.5466	4.4396	4.6183
Utilities	6.5651	3.2091	2.7465

1. Introduction
2. Data and Software
3. Methodology
4. Result
5. Possible Improvement

Possible Improvement

- *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 portfolio optimization to mitigate risks.*
- *Investigating advanced ensemble integration methods such as adaptive weighting to potentially enhance performance.*

Thanks for your listening!