Directional Stock Price Forecasting Based on Quantitative Value Investing Principles for Loss Averted Bogle-Head Investing using Various Machine Learning Algorithms

Agnij Moitra[a]

[a]*Amity International School, Saket, New Delhi, 110017, Delhi, India*

Abstract

Boglehead investing, founded on the principles of John C. Bogle is one of the classic time tested long term, low cost, and passive investment strategy. This paper uses various machine learning methods, and fundamental stock data in order to predict whether or not a stock would incur negative returns next year, and suggests a loss averted bogle-head strategy to invest in all stocks which are expected to not give negative returns over the next year. Results reveal that XGBoost, out of the 44 models trained, has the highest classification metrics for this task. Furthermore, this paper shall use various machine learning methods for exploratory data analysis, and SHAP values reveal that Net Income Margin, ROA, Gross Profit Margin and EBIT are some of the most important factors for this. Also, based on the SHAP values it is interesting to note that the current year has negligible contribution to the final prediction. Investors can use this as a heuristic guide for loss averted long term (1-year) stock portfolios.

Keywords: Stock market, Stock picking, Directional price forecasting, Machine learning
JEL: C38, C53, G17, G11, C45

1. Introduction

Traditional time-tested investment methods like Bogle-head strategies (Bogle, 1999, 2015a,b) have emphasized on passively investing in a diverse range of stocks, which quintessentially boils down to investing in low cost index funds (Larimore et al., 2011; Larimore, 2018; Lindauer et al., 2021; Berger, 2019). Most for the current machine learning based research (Liu and Long, 2020; Parray et al., 2020; Na and Kim, 2021; Kumbure et al., 2022) have primarily focused on using time-series based methods for short term trading which has empirically yielded favorable results, but is heavily dependent on the time-periods, *timing the market* (Bolton et al., 2013), and prone to overfitting and high variance since it is quite sensitive to the price-action patterns. Thereby, it is not reinforced by robust foundational and empirical analysis as concluded by (Sharpe, 1975; Roth and Xing, 1994; Jiang, 2003).

The contribution of this research is to use various machine learning algorithms and quantitative fundamental financial ratios and metrics for directional stock price forecasting for US stocks. Using fundamental financial metrics is supported by (Graham et al., 1934; Graham, 1949; Buffett, 1984; Klarman, 1991; Damodaran, 2011), although it should be noted that a major limitation of this is the fact that it cannot account for intangibles. Then use SHAP values to find the most relevant features for this task. This can be used in order to filter out stocks which are likely to yield negative returns, and use it as a loss-averted Bogle-head investment strategy.

Email address: [fullname][at]outlook[dot]com (Agnij Moitra)

2. Literature Review

Yan and Zheng (2017) created a large set of fundamental signals from financial statements and bootstrapped their way to investigating how data mining affected the anomalies based on fundamentals. They conclude that many fundamental signals survive data mining and are significant predictors of cross-sectional stock returns. Building on this, Hiransha et al. (2018) later showed that CNN gave the best results for predicting stock prices among MLP, RNN, and LSTM, using historical data from NSE (National Stock Exchange, India) and NYSE (New York Stock Exchange, USA). Their study demonstrated that neural networks significantly outperform traditional linear models such as ARIMA. Moreover, Huang et al. (2021) studied how three machine learning algorithms could be used in fundamental analysis for stock prediction and revealed that the best performing model is based on the Random Forest. Actually, the aggregated model outperformed both the baseline models and the benchmark DJIA index. A three-step feature engineering procedure was adopted to predict the direction of stock prices using a hybrid GA-XGBoost algorithm that enhances interpretability and also maintains a high level of accuracy within less computational cost compared to deep learning models. The emphasis is based on the idea that the feature-engineering process greatly improves the prediction performance. Further, Tsai et al. (2023) the use of machine learning methodologies for predicting stock returns in Taiwan financial markets using financial ratios. The authors found that the portfolios that are made up of only those top stocks selected by the models outperform the benchmark TW50 index, thus confirming the effectiveness of using the aligned financial ratios when making investment decisions with a mid- to long-term scope. On the same note, Zhao et al. (2023) applied machine learning models to examine ten categories of financial indicators at the Chinese stock exchange for predictive power. It has been found that these indicators are actually a high predictor of stock returns, and particularly the neural network models outperform linear ones. Nti et al. (2020) conducted a systematic review of more than 120 works related to the prediction of the stock market by machine learning, classified according to either technical, fundamental, or combined analysis, and reported that among widely used algorithms are support vector machines and artificial neural networks.

3. Methodology

3.1. Data Collection and pre-processing

For the dependent variables, 25 financial ratios and metrics for all available US stocks in tikr.com were collected from from 2016 to 2023, this was combined with 10 economic factors—inflation, AAA bond rate, and 10 years T-bills et cetera—for the corresponding years. The returns for each stock and whether or not it yielded positive return were retrived using the Alpha Vantage API. The data collection process for similar to that of Moitra (2023) (See Section Appendix C for discription of the feature collected). The data collected underwent the various pre-processing to ensure consistency. To deal with missing values K-Nearest Neighbors (KNN) imputer (Troyanskaya et al., 2001) was used, which estimates missing values X_{ij} by averaging the j-th feature values of the k-nearest neighbors of sample i:

$$X_{ij} = \frac{1}{|N_{ij}|} \sum_{l \in N_{ij}} X_{lj}$$

Next outliers were removed Local Outlier Factor (Breunig et al., 2000), since it may reduce the skewness of the data thus making the machine learning model more generalized. The Local Outlier Factor (LOF) of a point x is defined as:

$$\text{LOF}(x) = \frac{\frac{1}{|N_k(x)|} \sum_{o \in N_k(x)} \text{lrd}(o)}{\text{lrd}(x)}$$

where $N_k(x)$ is the set of k-nearest neighbors of x, and $\text{lrd}(x)$ is the local reachability density of x, measuring how isolated x is compared to its neighbors. In practice this was done using Scikit-Learn with 20 N-Neighbours with a Minkowski distance of 2.

3.2. Machine learning models

After all the preprocessing the dataset had around 25,000 samples. And for the directionality prediction all the 41 available classifiers from Scikit-Learn (Pedregosa et al., 2011), and gradient boosting methods—XGBoost, LightGBM & CatBoost (Chen and Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018)—were used.

Gradient boosting is a machine learning technique to iteratively train weak learners, usually decision trees, to minimize the residual errors. Wherein the initial model is $F_0(x) = \arg\min \sum_{i=1}^{n} \mathcal{L}(y_i, \gamma)$, where $\mathcal{L}(y, \gamma)$ is the loss function minimized by a constant γ (in practice this is the arithmetic mean). For each iteration n, the pseudo-residuals are calculated, i.e the gradient of the loss function with respect to the model's predictions, such that $r_i^{(n)} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{n-1}}$. A weak learner $h_m(x)$ is trained on these residuals and the optimal learning rate $\alpha_n = \arg\min_\alpha \sum_{i=1}^{N} L(y_i, F_{n-1}(x_i) + \alpha h_n(x_i))$, and then the model is updated as:

$$F_n(x) = F_{n-1}(x) + \alpha_n h_n(x)$$

And after N iterations the final model is $F_M(x) = F_0(x) + \sum_{m=1}^{M} \alpha_m h_m(x)$, where all the weak learners are summed up to produce a strong predictive model.

4. Results & Discussion

4.1. Exploratory Data Analysis

The t-SNE plot (Figure 1) of the dependent shows that the data has some degree of clusterization. Although the stocks with negative return generally form a dense cluster in the bottom center, yet the positive and negative return clusters are overlapping each other, which makes it a non-trivial classification task.

The correlation heatmap of financial variables (Figure 2) illustrates numerous relationships among metrics: for all liquidity ratios and for all profitability metrics, including ROA, and EBIT, the correlations are high and positive, such that the improvement of one metric is usually accompanied by the improvement of another in its category. Correlations are also negative among leverage ratios, such as Debt to Equity, and the various measures of profitability, which means that higher leverage is associated with lesser profitability. The separate clusters of correlated variables, such as those between liquidity ratios and the profitability metrics, give further evidence of the intertwined character of the indicators of financial health. In addition, there are market and economic indicators, such as the bond yields and the price of gold, that indicate correlations due to overall economic conditions.

4.2. Preliminary Benchmarking

For the preliminary testing, all models from Scikit-Learn, XGBoost, LightGBM, and CatBoost. And it was found that gradient boosting methods, and bagging (Random Forest and Extra Trees) had the highest accuracy, ROC AUC, and F1 score. (See Table B.4 for the entire unabridged results) Furthermore, other versions with polynomial transformation, spline transformation, and with PCA (Tipping and Bishop, 1999), 1 layer Artificial Neural Network and Restricted Boltzmann Machine Rumelhart et al. (1986); Hinton et al. (2006); Tieleman (2008), which were implemented using PyTorch (Paszke et al., 2019), and (with same number of components/hidden dimensions as the number of features as prescribed by Tannor and Rokach (2019) for feature extraction) were tested, but these feature extraction methods did not result in a statistically significant improvement in any of the metrics being considered, as per paired single tail t-test (p-value > 0.6).

4.3. Hyper-parameter Tuning

The top-5 models from the initial benchmarking were then hyper-parameter tuned using Scikit-Learn's Randomized Grid Search CV with 100 iterations and 5-flod cross validation, and it turns out that XGBoost was the best model for this, with accuracy—0.816 ± 0.012, ROC AUC—0.816 ± 0.007, F1 score—0.815 ± 0.016, and was trained with leave-one-out cross validator on 70% of the dataset.

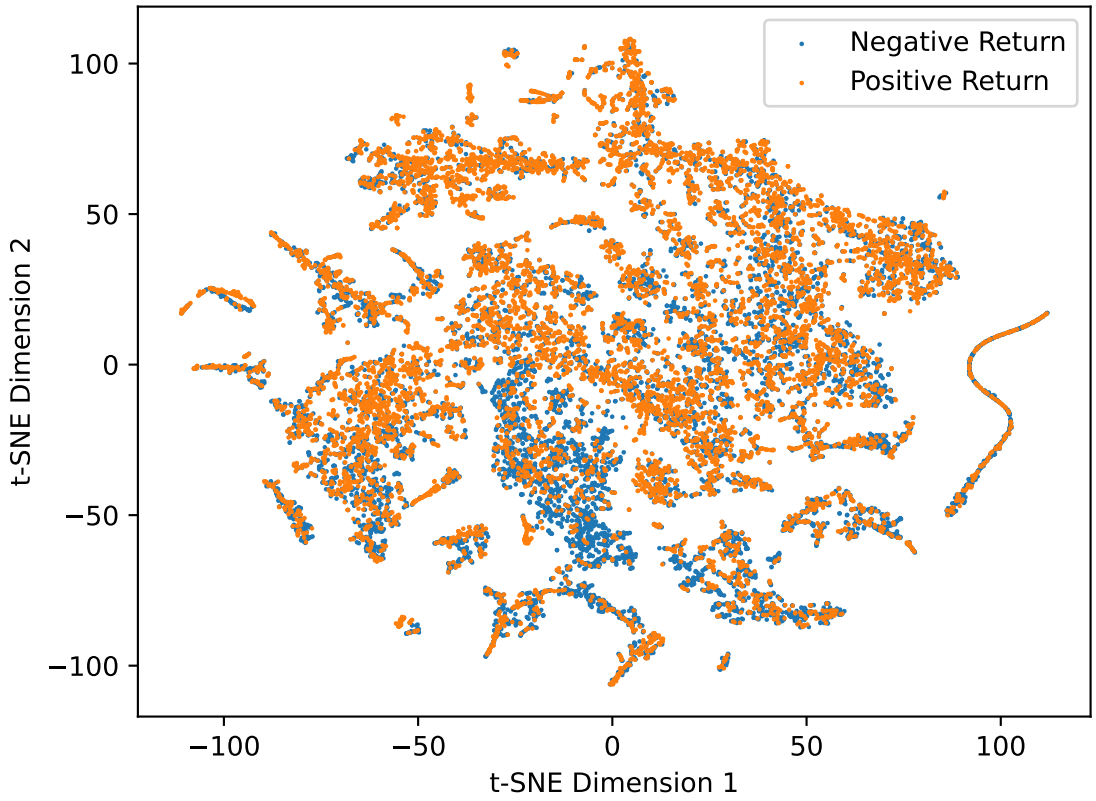

Figure 1: t-SNE Visualization (Positive return: orange point, negative return: blue point)

Table 1: Top-5 Model Comparison to predict whether or not a company would yield positive return on investment over the next year

	Accuracy	**ROC AUC**	**F1**
CatBoost	0.73	0.71	0.72
Gradient Boostings	0.72	0.71	0.72
LGBM	0.72	0.70	0.71
Histogram Gradient Boosting	0.72	0.70	0.71
XGBoost	0.70	0.70	0.70

Table 2: XGBoost hyperparameters

	Description
`subsample` (0.8)	Fraction of training data used per tree (0.8 means 80%).
`n_estimators` (2500)	Number of trees in the model
`max_depth` (14)	Maximum depth of each tree
`learning_rate` (0.01)	Shrinks contribution of each tree
`gamma` (0.2)	Minimum loss reduction for further partitioning
`colsample_bytree` (0.85)	Fraction of features used per tree

4.4. Comparison with low-cost index funds

Whilst accuracy, AUC ROC, and F1 score are useful metrics, back-testing the model and comparing it to other alternatives is equally if not more important. Thereby, the final XGBoost model was used to select stocks, which is expected to yield positive return over the next year (i.e from January 1st Year to

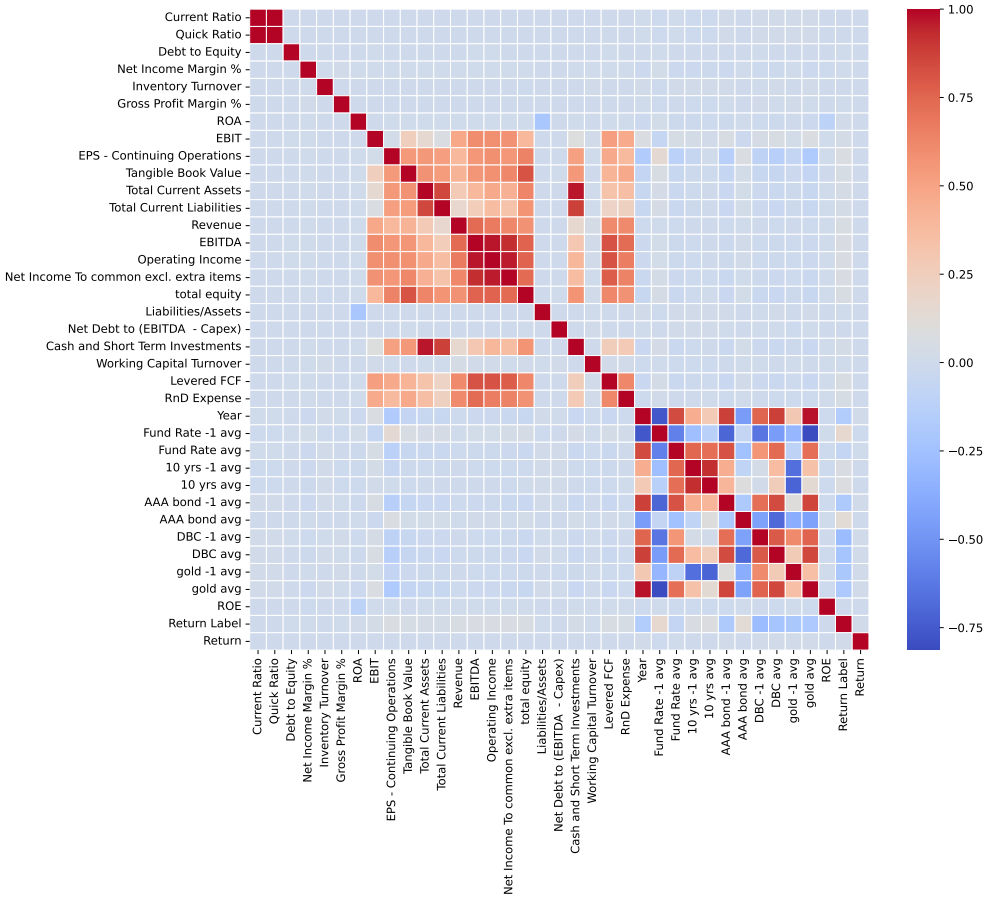

Figure 2: Correlation Heatmap of the dependent variables and next year's return (return on investment) label and return (return on investment)

January 1st Year + 1), from Standard and Poor's 500 (SPY), Dow Jones Industrial Average (DJI), NASDAQ Composite (IXIC), and Russell 2000 (RUT). (See Table B.5 for the actual values) There is a sight statistically significant improvement if this machine learning algorithm is used with SPY and DJI, although the absolute improvement is rather significant (Table B.5). And it didn't do quite as well on NASDAQ Composite this can be attributed to the fact that NASDAQ is skewed towards *growth stocks* (Chan and Lakonishok, 2004; Cronqvist et al., 2015) with very high price to earning ratios, whose financial ratios may not truly reflect the intangibles of their business, unlike a *value stock*. But there is significant statistical evidence to use this method on small cap stocks as highlighted by the p-value for return on investment of Russell 2000. This is consistent with the fact that small cap stocks and the *less attractive businesses*, which are often over-looked by the main-stream investment analysts and the media, are a good place to find bargains and mispriced opportunities. This theory is reinforced by many previous works Greenblatt (2010); Marks (2011); Oxman et al. (2011); Carlisle (2017) and cult-classics like (Graham et al., 1934; Graham, 1949; Buffett, 1984; Klarman, 1991).

4.5. Feature Importance & Explainability

Finding which features play a key role in making predictions is quite necessary to gain insights about the data and verify basic facts to see if they are consistent with previously well-received ideas. Which this research does SHAP (Lundberg and Lee, 2017) values. The SHAP heatmap (Figure 4) for the fine-tuned XGBoost model reveals that Net income margin %, ROA, the sector of the company, EBIT, EPS are some

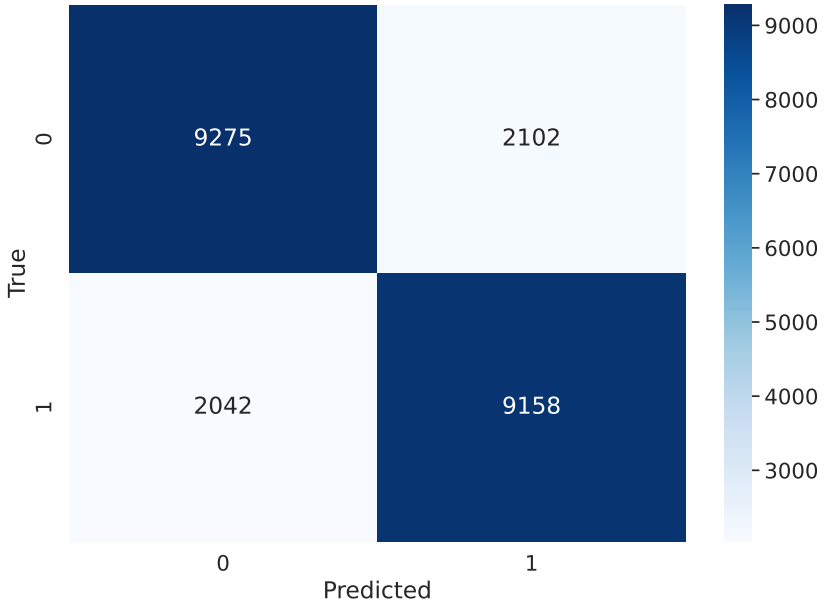

Figure 3: Confusion Matrix of XGBoost (Tuned)

Table 3: p-values for paired single-tailed t-tests for XGBoost vs Index Funds (ROI: Return of investment (alternate: greater), σ: Standard deviation (alternate: less), Sharpe Ratio (alternate: greater), Sortino (alternate: greater), ρ: Market Correlation (alternate: less))

	SPY	DJI	IXIC	RUT
ROI	0.092	0.105	0.489	0.040
σ	0.406	0.443	0.696	0.116
Sharpe	0.112	0.166	0.488	0.194
Sortino	0.102	0.166	0.244	0.158
ρ	5.7e-11	7.7e-07	0.003	0.07

of the most important financial information for predicting whether or or a stock would yield positive return on investment. Furthermore, macroeconomic factors like inflation (measured by Invesco DB Commodity Index i.e DBC a significant factor to determine operating costs and other expenses), AAA bond rate and federal funds effective rate. Net income margin %, EBIT indeed one of the important metrics considered by value investors as noted by Graham et al. (1937). Also gross-profit margin % (a measure of a company's *competitive moat*), return of equity, return on assets as per (Buffett and Cunningham, 2001; Buffett and Clark, 2008). Although it is interesting to note that current ratio, current assets, liabilities/assets, tangible book value, quick ratio which were also quintessential part of what Benjamin Graham described Net-nets and *cigar-butt* stocks, but such mis-priced liquidation/acquisition opportunities are rather rare in today's markets and accordingly their low SHAP values can be justified. Furthermore, since the model only tried to predict the companies performance over a short period of time, i.e 1 year, accordingly short term holdings and levered free cash-flow did have a significant SHAP value. Moreover, net income, operating income and EBITDA have a noticeable skew towards reducing the returns, thus these metrics can potentially be used for loss aversion; and, similarly 10 years T-bills, AAA bond rate, gold returns also have a skew towards reducing the returns, this can be attributed to the fact that investors do pay attention to relatively *safe* and less volatile alternatives before investing in stocks, and 10 years T-bills is especially important for financial and banking companies. Apart from this, it is interesting to note that EBIT has a higher SHAP value that

EBITDA, as it does not account for depreciation and amortization. As an anecdote, The use of EBITDA was once famously critiqued by Charlie Munger in one of Berkshire Hathways' annual share holder's meeting, and Chullen et al. (2015) also reached a similar conclusion that EBITDA is less important than EBIT.

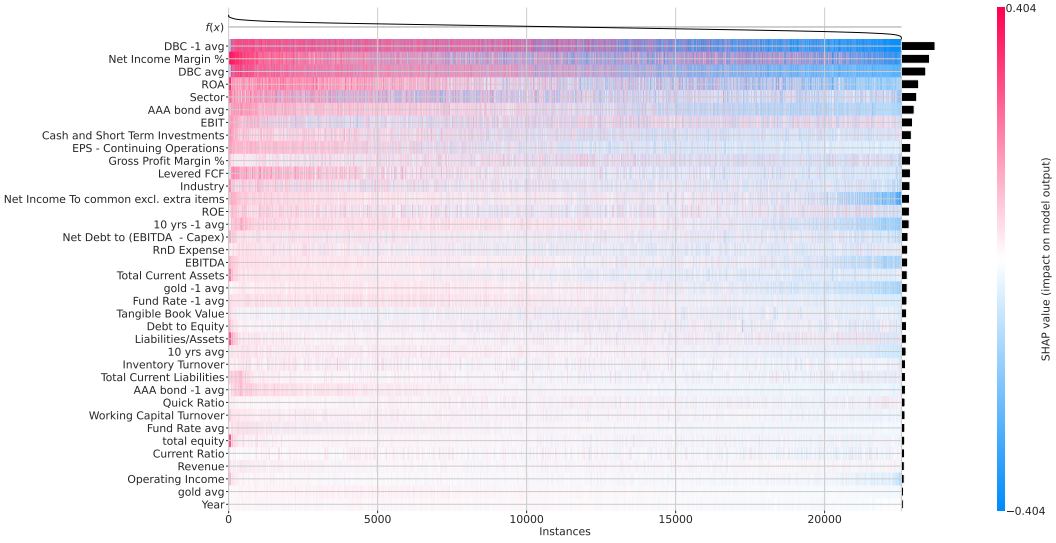

Figure 4: SHAP value heatmap of XGBoost (Hyper-parameter tuned)

4.6. Directly predicting returns

As found in the above sub-sections, machine learning models are fairly accurate enough to classify whether or not a company would have positive or negative return on investment over the next year, based on the current year's financial information. Thereby, it is natural to hypothesize that regression models would be used to directly predict the absolute returns and then perhaps invest in the top-N companies. But empirical testing shows that even at the best case scenario histogram gradient boosting regressor had a root mean squared error of 81.94. And as per the residual plot (Figure 5) there is a considerable variance of the errors, and it skewed towards negative residual errors which implies that the predictions are *overly optimistic* and inaccurate when compared to the actual returns. Thereby is it not feasible to use machine learning models to make return on investment predictions directly. (See Table B.6 for the complete results)

4.7. Limitations & Considerations

The following are a few identified limitations and considerations of this research:

1. Due to data unavailability this research could not use international stocks, which may have added additional insights to the machine learning model. Furthermore, this research may not be as accurate whilest dealing with companies whose significant income is determined by sales in international markets.
2. Even though, leave one out cross validator (on 70% on the dataset randomly sampled and stratified) was used in order to minimize the skew towards training data. Backtesting should not be considered as a guarantee for future returns, since the model may still have biases from the training data.
3. Investors should be mindful of the fact that even though, the machine learning model gives a great improvement in percentage terms, it only holds a slight statistical significance for large cap blue chip stocks. Which are already in the radar of financial analysts, so there's less room for mis-priced

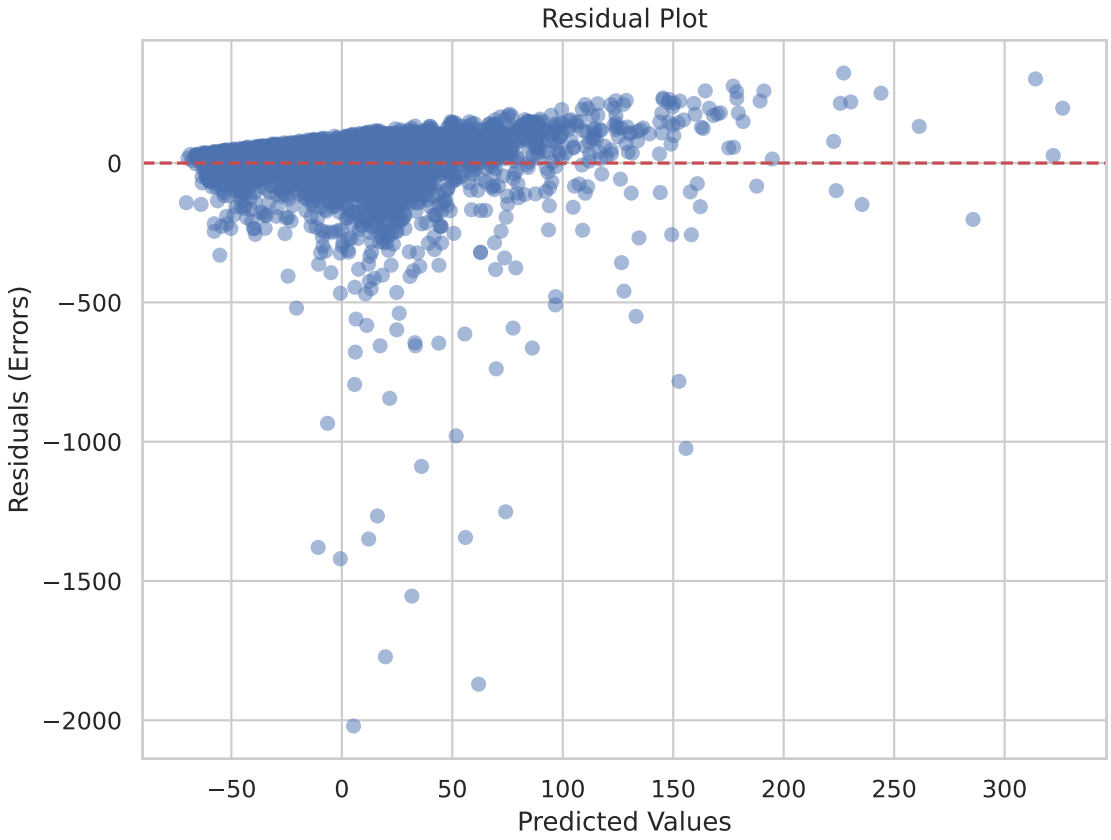

Figure 5: Residual plot for the histogram gradient boosting regressor

opportunities for a value investor. But it did have a statistically significant improvement for small cap and lesser known companies, and reducing the overall market correlation ρ.

4. Since the machine learning model was trained to deal with one calendar year to the next, it should theoretically yield optimal returns when it is used like that.
5. In theory, grid search CV is a more exhaustive and robust method to find the best hyper-parameters than randomized grid search CV. But this research could not use it due to computational resource constrains.
6. This research should not be used to directly invest in individual stocks, as is the case with many of the other conventional methods like Magic Formula method (Greenblatt, 2010) and the Aquirer's Multiple (Carlisle, 2017), but rather filter a portfolio of stocks to invest in and research regarding the intangibles of that company, which this research cannot account for.

5. Conclusion

This paper uses multiple machine learning methods to predict whether or not a company would yield positive return on investment over the next calender year—wherein fined tuned XGBoost had the highest accuracy. And it turns out that Net income margin %, ROA, EBIT, EPS are some of the most important factors for this, and inflation, 10 years T-bills, AAA bond rate are significant macroeconomic factors to consider. Also, it is found that because of high root mean squared error and variance in the predicting the

returns directly, investors are better-off with a portfolio of loss-averted stocks, than individually finding the companies which would yield the highest return on investment. Future works may include more data from various countries and spanning multiple years, and adding quarterly financial data as well for more informed predictions, encoding intangibles of a business (which can be done using a language model (Vaswani et al., 2017) or traditional methods like Td-Idf (Zhang et al., 2011), Word2Vec (Mikolov et al., 2013)). Also as an alternative to the current gradient boosting methods, MSBoost (Moitra, 2024), could be used but it does require quite a lot of computational resources to be trained on a large dataset with many features.

Appendix A. Data Availability Section

Both TIKR's and Alpha Vantage's terms of use policy Section 3.1 & Section 2.a only allows personal use of the financial data, thereby the data used for this research cannot be made publicly available.

Appendix B. Additional Results

This section contains the full-length tables for classification results (TableB.4), comparison with low-cost index funds (Table B.5), and regression task results (Table B.6).

Table B.4: Model Comparison to predict whether or not a company would yield positive return on investment over the next year

	Accuracy	**Balanced Accuracy**	**ROC AUC**	**F1 Score**
CatBoostClassifier	0.73	0.71	0.71	0.72
GradientBoostingClassifier	0.72	0.71	0.71	0.72
LGBMClassifier	0.72	0.70	0.70	0.71
HistGradientBoostingClassifier	0.72	0.70	0.70	0.71
XGBClassifier	0.72	0.70	0.70	0.71
ExtraTreesClassifier	0.71	0.70	0.70	0.71
RandomForestClassifier	0.71	0.70	0.70	0.70
MLPClassifier	0.70	0.69	0.69	0.70
AdaBoostClassifier	0.70	0.69	0.69	0.70
BaggingClassifier	0.69	0.68	0.68	0.69
GaussianProcessClassifier	0.70	0.68	0.68	0.69
SVC	0.70	0.68	0.68	0.69
NuSVC	0.69	0.68	0.68	0.69
LogisticRegression	0.70	0.67	0.67	0.69
LogisticRegressionCV	0.70	0.67	0.67	0.69
KNeighborsClassifier	0.69	0.67	0.67	0.68
CalibratedClassifierCV	0.70	0.67	0.67	0.68
SGDClassifier	0.69	0.67	0.67	0.68
LinearSVC	0.69	0.67	0.67	0.68
BernoulliNB	0.69	0.66	0.66	0.67
RidgeClassifierCV	0.68	0.65	0.65	0.66
NearestCentroid	0.68	0.65	0.65	0.66
RidgeClassifier	0.68	0.65	0.65	0.66
LinearDiscriminantAnalysis	0.68	0.65	0.65	0.66
Perceptron	0.66	0.64	0.64	0.65
PassiveAggressiveClassifier	0.64	0.63	0.63	0.64
LabelSpreading	0.64	0.63	0.63	0.64
LabelPropagation	0.64	0.63	0.63	0.64
DecisionTreeClassifier	0.62	0.62	0.62	0.62
ExtraTreeClassifier	0.62	0.61	0.61	0.62
QuadraticDiscriminantAnalysis	0.56	0.59	0.59	0.54
GaussianNB	0.51	0.56	0.56	0.47
DummyClassifier	0.57	0.50	0.50	0.41

Appendix C. Discription of the features used

The following contains brief descriptions of the dependent variable collected (GPT 3.5 was used to make this):

Table B.5: Comparison with other low-cost index funds (ROI: Return on investment, σ: Standard deviation, ρ: Market Correlation)

	Metric	2016	2017	2018	2019	2020	2021	2022	2023	Avg.
Standard and Poor's 500 (SPY)	ROI (%)	11.96	21.83	-4.38	31.22	18.40	28.71	-18.11	11.15	12.60
	σ	10.62	6.69	11.51	12.09	19.91	13.23	23.51	16.43	14.25
	Sharpe	0.98	3.08	-0.36	2.26	0.92	2.11	-0.82	0.68	1.11
	Sortino	1.52	4.94	-0.54	3.48	1.40	3.30	-1.20	1.02	1.74
	ρ	0.98	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.99
XGBoost on SPY stocks	ROI (%)	20.00	35.42	-0.45	51.90	29.58	47.36	0.68	17.57	26.36
	σ	10.33	6.36	11.87	10.48	18.33	13.13	22.40	16.04	13.62
	Sharpe	1.81	5.56	-0.18	4.46	1.50	3.69	-0.11	1.10	2.39
	Sortino	1.86	16.80	0.01	9.96	1.33	7.36	0.03	0.76	5.34
	ρ	0.74	0.80	0.83	0.78	0.84	0.79	0.80	0.77	0.79
Dow Jones Industrial Average (DJI)	ROI (%)	13.42	25.08	-3.48	22.34	9.72	18.73	-8.78	6.95	10.50
	σ (%)	10.22	6.78	10.94	12.01	20.85	12.15	21.71	15.36	13.75
	Sharpe	1.14	3.50	-0.31	1.79	0.46	1.54	-0.41	0.45	1.02
	Sortino	1.77	5.60	-0.47	2.76	0.70	2.41	-0.60	0.68	1.61
	ρ	0.97	0.98	0.98	0.98	0.98	0.98	0.97	0.98	0.98
XGBoost on DJI stocks	ROI (%)	21.89	41.43	0.29	36.58	16.90	31.76	-1.63	11.21	21.03
	σ	10.09	6.70	10.61	11.78	20.17	11.97	21.46	14.28	13.38
	Sharpe	2.13	5.96	-0.11	3.19	0.68	2.58	-0.06	0.65	2.05
	Sortino	2.44	22.08	-0.03	5.44	0.39	4.10	-0.01	0.34	4.92
	rho	0.73	0.75	0.83	0.84	0.84	0.80	0.91	0.84	0.82
NASDAQ Composite (IXIC)	ROI (%)	7.5	28.2	-3.9	35.2	43.6	21.4	-32.5	31.0	16.31
	σ (%)	12.4	13.1	16.7	12.9	20.5	17.3	23.7	19.0	16.95
	Sharpe	0.6	1.9	-0.2	2.2	1.8	1.2	-1.4	1.5	0.95
	Sortino	0.8	2.6	-0.3	3.0	2.7	1.7	-1.8	2.0	1.34
	ρ	0.95	0.92	0.94	0.96	0.97	0.95	0.93	0.94	0.94
XGBoost on IXIC stocks	ROI (%)	5.75	31.34	-5.33	35.18	42.56	21.92	-27.99	29.71	14.78
	σ	10.86	16.53	16.59	15.62	20.04	19.44	23.73	21.21	18.0
	Sharpe	0.47	2.00	-0.29	2.40	1.87	1.13	-1.30	1.47	0.9
	Sortino	0.11	3.97	-0.02	5.27	2.99	1.39	0.08	2.10	1.97
	ρ	0.86	0.92	0.93	0.87	0.90	0.92	0.95	0.90	0.91
Russell 2000 (RUT)	ROI (%)	21.3	14.6	-11.0	25.5	18.4	14.8	-20.5	13.5	9.57
	σ (%)	16.1	13.0	16.0	15.2	22.5	19.8	22.7	19.0	18.04
	Sharpe	1.3	1.0	-0.7	1.5	0.8	0.7	-1.0	0.7	0.54
	Sortino	1.8	1.4	-0.9	2.1	1.2	1.0	-1.3	1.0	0.79
	ρ	0.87	0.85	0.88	0.90	0.91	0.89	0.86	0.88	0.88
XGBoost on RUT stocks	ROI (%)	45.65	48.87	24.97	34.83	34.73	9.34	-3.84	13.42	27.79
	σ	33.10	2.86	7.49	24.10	19.43	13.13	0.61	-1.96	12.35
	Sharpe	2.06	2.01	-0.51	1.86	1.87	0.61	-1.19	1.49	0.96
	Sortino	3.64	1.97	-0.24	6.85	0.49	0.11	-1.14	3.13	1.67
	ρ	1.18	0.78	0.36	1.17	0.27	0.85	0.51	0.24	0.67

- Current Ratio: Measure of a company's liquidity and ability to meet short-term obligations. Useful for assessing financial stability.
- Quick Ratio: Similar to the current ratio but focuses on a company's most liquid assets, providing insight into its short-term financial health.
- Debt to Equity: Indicates the level of financial leverage and potential risk. Helps assess a company's ability to handle debt and financial health.
- Net Income Margin: Indicates the percentage of revenue that remains as profit after all expenses are deducted.
- Inventory Turnover: Reflects how efficiently a company manages its inventory, which can impact cash flow and profitability.
- Gross Profit Margin: Indicates the percentage of revenue that remains as profit after all expenses are deducted.

Table B.6: Model comparison for regression task (Predicting the absolute returns)

	Adjusted R-Squared	R-Squared	RMSE
HistGradientBoostingRegressor	0.07	0.07	81.94
ExtraTreesRegressor	0.06	0.06	82.48
GradientBoostingRegressor	0.06	0.06	82.59
MLPRegressor	0.05	0.05	82.78
Lasso	0.05	0.05	82.84
LassoLars	0.05	0.05	82.84
LassoLarsCV	0.05	0.05	82.85
LassoCV	0.05	0.05	82.85
LassoLarsIC	0.05	0.05	82.90
LGBMRegressor	0.05	0.05	82.90
OrthogonalMatchingPursuit	0.05	0.05	82.91
ARDRegression	0.05	0.05	82.94
BayesianRidge	0.05	0.05	83.03
RidgeCV	0.04	0.05	83.09
Ridge	0.04	0.05	83.10
TransformedTargetRegressor	0.04	0.05	83.10
LinearRegression	0.04	0.05	83.10
ElasticNetCV	0.04	0.04	83.22
LarsCV	0.04	0.04	83.23
OrthogonalMatchingPursuitCV	0.04	0.04	83.28
KernelRidge	0.04	0.04	83.45
NuSVR	0.04	0.04	83.45
ElasticNet	0.03	0.04	83.50
SVR	0.03	0.04	83.58
TweedieRegressor	0.03	0.03	83.75
CatBoostRegressor	0.03	0.03	83.75
RandomForestRegressor	0.02	0.03	84.01
PLSRegression	0.01	0.02	84.43
DummyRegressor	-0.00	-0.00	85.13
QuantileRegressor	-0.01	-0.01	85.51
KNeighborsRegressor	-0.03	-0.03	86.40
XGBRegressor	-0.04	-0.03	86.51
LinearSVR	-0.06	-0.06	87.49
BaggingRegressor	-0.08	-0.08	88.40
HuberRegressor	-0.19	-0.19	92.72
PassiveAggressiveRegressor	-0.20	-0.19	93.02
Lars	-0.20	-0.20	93.08
DecisionTreeRegressor	-1.07	-1.07	122.36
ExtraTreeRegressor	-1.32	-1.31	129.45
AdaBoostRegressor	-4.65	-4.63	202.06
TheilSenRegressor	-4.95	-4.93	207.29
GaussianProcessRegressor	-8.50	-8.47	261.97
RANSACRegressor	-111.01	-110.61	899.38

- ROA (Return on Assets): Provides insights into how efficiently a company utilizes its assets to generate profit, important for assessing overall performance.
- EBIT (Earnings Before Interest and Taxes): Reflects a company's operating performance, excluding interest and tax expenses, providing insight into core profitability.
- EPS - Continuing Operations (Earnings per Share from Continuing Operations): Measures earnings available to common shareholders and is essential for assessing shareholder value.
- Tangible Book Value: Represents the net value of a company's tangible assets, which can be useful for evaluating a company's intrinsic value.
- Total Current Assets: An indicator of a company's short-term liquidity and ability to meet immediate financial obligations.
- Total Current Liabilities: Shows a company's short-term financial obligations and potential liquidity risk.
- Revenue: Reflects a company's total sales and is a fundamental metric for assessing business performance.

- EBITDA (Earnings Before Interest, Taxes, Depreciation, and Amortization): A measure of operating performance that excludes non-cash expenses.
- Operating Income: Indicates a company's profitability from its core operations, excluding non-operating income and expenses.
- Net Income To Common excl. extra items: Reflects the net income available to common shareholders, excluding extraordinary items, important for assessing shareholder value.
- Total Equity: Shows a company's net assets, which can be important for understanding its financial health and leverage.
- Liabilities/Assets: A ratio reflecting a company's financial leverage and potential risk.
- Net Debt to (EBITDA - Capex): Helps evaluate a company's ability to manage debt in relation to its operating cash flow.
- Cash and Short Term Investments: Indicates a company's cash resources, which can be crucial for assessing liquidity and financial flexibility.
- Working Capital Turnover: Measures how efficiently a company uses its working capital, which can impact cash flow and profitability.
- Levered FCF (Levered Free Cash Flow): Measures the cash a company generates after expenses and debt service, important for assessing financial health.
- R&D Expense (Research and Development Expenses): Reflects investments in innovation and product development, which can influence future competitiveness.
- Year: Specifies the year of the data, important for tracking historical trends and making time-series analyses.
- Fund Rate -1 Change (Average Fund Rate for the Previous Period): Provides insights into the interest rate environment, which can impact investment decisions, (source: Statista).
- Fund Rate avg (Average Fund Rate): Reflects the current average fund rate, useful for understanding the prevailing interest rate conditions, (Source: Statista).
- 10 yrs -1 Change (Average 10-Year Bond Rate for the Previous Period): Provides insights into the interest rate environment, particularly for longer-term bonds, (Source: `https://markets.ft.com/data/bonds/tearsheet/charts?s=US10YT`).
- 10 yrs Open (Average 10-Year Bond Rate): Reflects the current average yield on 10-year bonds, which can influence investment decisions, (Source: `https://markets.ft.com/data/bonds/tearsheet/charts?s=US10YT`).
- AAA bond -1 Open (Average AAA Bond Rate for the Previous Period): Indicates the yield on AAA-rated bonds from the previous period, offering insights into the bond market, (Source: `https://fred.stlouisfed.org/series/AAA`).
- AAA bond avg (Average AAA Bond Rate): Shows the current yield on AAA-rated bonds, which can be relevant for assessing the fixed income market, (Source: `https://fred.stlouisfed.org/series/AAA`).
- DBC -1 Change (Average Commodity Index for the Previous Period): Indicates the average commodity index from the previous period, which is relevant for commodity market analysis, (Source: `https://www.averageannualreturn.com/dbc/(year).html`).

- DBC Open (Average Commodity Index): Shows the current average commodity index, valuable for assessing the commodity market's current state, (Source: `https://www.averageannualreturn.com/dbc/(year).html`).
- Gold -1 Change (Average Gold Price for the Previous Period per Oz): Reflects the average gold price from the previous period, which is essential for gold market analysis, (Source: goldprice.org).
- Gold Open (Average Gold Price per Oz): Specifies the current average gold price, valuable for assessing the current state of the gold market, (Source: `https://goldprice.org/gold-price-chart.html`).

References

Berger, R., 2019. Retire Before Mom and Dad: The Simple Numbers Behind A Lifetime of Financial Freedom. BiggerPockets Publishing.

Bogle, J.C., 1999. Common sense on mutual funds: New imperatives for the intelligent investor. John Wiley & Sons.

Bogle, J.C., 2015a. Bogle on mutual funds: New perspectives for the intelligent investor. John Wiley & Sons.

Bogle, J.C., 2015b. John Bogle on investing: the first 50 years. John Wiley & Sons.

Bolton, P., Chen, H., Wang, N., 2013. Market timing, investment, and risk management. Journal of Financial Economics 109, 40–62.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. Lof: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Association for Computing Machinery. pp. 93–104.

Buffett, M., Clark, D., 2008. Warren Buffett and the interpretation of financial statements: The search for the company with a durable competitive advantage. Simon and Schuster.

Buffett, W., Cunningham, L.A., 2001. The essays of Warren Buffett: lessons for corporate America. HeinOnline.

Buffett, W.E., 1984. The superinvestors of graham-and-doddsville. Hermes, Columbia Business School , 4–15.

Carlisle, T.E., 2017. The Acquirer's Multiple: How the Billionaire Contrarians of Deep Value Beat the Market. Ballymore Publishing.

Chan, L.K., Lakonishok, J., 2004. Value and growth investing: Review and update. Financial Analysts Journal 60, 71–86.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd AMC SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Chullen, A., Kaltenbrunner, H., Schwetzler, B., 2015. Does consistency improve accuracy in multiple—based valuation? Journal of Business Economics 85, 635–662.

Cronqvist, H., Siegel, S., Yu, F., 2015. Value versus growth investing: Why do different investors have different styles? Journal of Financial Economics 117, 333–349.

Damodaran, A., 2011. The little book of valuation: How to value a company, pick a stock, and profit. Wiley.

Graham, B., 1949. The Intelligent Investor: A Book of Practical Counsel. Harper & Brothers.

Graham, B., Dodd, D.L.F., Cottle, S., et al., 1934. Security analysis. volume 452. McGraw-Hill New York.

Graham, B., McGolrick, C., Meredith, S.B., 1937. The interpretation of financial statements. volume 4. Harper York.

Greenblatt, J., 2010. The little book that still beats the market. John Wiley & Sons.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural computation 18, 1527–1554.

Hiransha, M., Gopalakrishnan, E.A., Menon, V.K., Soman, K., 2018. Nse stock market prediction using deep-learning models. Procedia computer science 132, 1351–1362.

Huang, Y., Capretz, L.F., Ho, D., 2021. Machine learning for stock prediction based on fundamental analysis, in: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE. pp. 01–10.

Jiang, W., 2003. A nonparametric test of market timing. Journal of Empirical Finance 10, 399–425.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30.

Klarman, S.A., 1991. Margin of Safety: Risk-Averse Value Investing Strategies for the Thoughtful Investor. Harper Business.

Kumbure, M.M., Lohrmann, C., Luukka, P., Porras, J., 2022. Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications 197, 116659.

Larimore, T., 2018. The Bogleheads' Guide to the Three-fund Portfolio: How a Simple Portfolio of Three Total Market Index Funds Outperforms Most Investors with Less Risk. John Wiley & Sons.

Larimore, T., Lindauer, M., Ferri, R.A., Dogu, L.F., 2011. The Bogleheads' Guide to Retirement Planning. John Wiley & Sons.

Lindauer, M., Larimore, T., LeBoeuf, M., 2021. The Bogleheads' Guide to Investing. John Wiley & Sons.

Liu, H., Long, Z., 2020. An improved deep learning model for predicting stock market price time series. Digital Signal Processing 102, 102741.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

Marks, H., 2011. The most important thing: uncommon sense for the thoughtful investor. Columbia University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26.

Moitra, A., 2023. Machine masquerades a poet: Using unsupervised t5 transformer for semantic style transformation in poetry generation, in: International Conference on Applied Informatics, Cham: Springer Nature Switzerland. pp. 403–418.
Moitra, A., 2024. Msboost: Using model selection with multiple base estimators for gradient boosting. HAL open science, Preprint doi:10.13140/RG.2.2.16729.33122.
Na, H., Kim, S., 2021. Predicting stock prices based on informed traders' activities using deep neural networks. Economics Letters 204, 109917.
Nti, I.K., Adekoya, A.F., Weyori, B.A., 2020. A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review 53, 3007–3057.
Oxman, J., Mohanty, S., Carlisle, T.E., 2011. Deep value investing and unexplained returns, in: Midwest Finance Association 2012 Annual Meetings Paper.
Parray, I.R., Khurana, S.S., Kumar, M., Altalbe, A.A., 2020. Time series data analysis of stock price movement using machine learning techniques. Soft Computing 24, 16509–16517.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Resesearch 12, 2825–2830.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems 31.
Roth, A.E., Xing, X., 1994. Jumping the gun: Imperfections and institutions related to the timing of market transactions. The American Economic Review , 992–1044.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. nature 323, 533–536.
Sharpe, W.F., 1975. Likely gains from market timing. Financial Analysts Journal 31, 60–69.
Tannor, P., Rokach, L., 2019. Augboost: Gradient boosting enhanced with step-wise feature augmentation, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization. pp. 3555–3561. doi:10.24963/ijcai.2019/493.
Tieleman, T., 2008. Training restricted boltzmann machines using approximations to the likelihood gradient, in: Proceedings of the 25th international conference on Machine learning, pp. 1064–1071.
Tipping, M.E., Bishop, C.M., 1999. Mixtures of probabilistic principal component analyzers. Neural computation 11, 443–482.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for dna microarrays. Bioinformatics 17, 520–525.
Tsai, P.F., Gao, C.H., Yuan, S.M., 2023. Stock selection using machine learning based on financial ratios. Mathematics 11, 4758.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.
Yan, X., Zheng, L., 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. The Review of Financial Studies 30, 1382–1423.
Zhang, W., Yoshida, T., Tang, X., 2011. A comparative study of tf* idf, lsi and multi-words for text classification. Expert systems with applications 38, 2758–2765.
Zhao, C., Yuan, X., Long, J., Jin, L., Guan, B., 2023. Financial indicators analysis using machine learning: Evidence from chinese stock market. Finance Research Letters 58, 104590.